MACRO: Enhancing Multilingual Counterfactual Explanations through Alignment-as-Preference Optimization

Anonymous ACL submission

Abstract

Self-generated counterfactual explanations (SCEs) are minimally modified inputs (*minimality*) generated by large language models (LLMs) that flip their own predictions (*validity*), offering a causally grounded approach to unraveling black-box LLM behavior. Yet extending them beyond English remains challenging: existing methods struggle to produce valid SCEs in non-dominant languages, and a persistent trade-off between *validity* and *minimality* undermines explanation quality. We introduce MACRO[1], a preference alignment framework that applies Direct Preference Optimization (DPO) to multilingual SCE generation, using a composite scoring function to construct preference pairs that effectively translate the trade-off into measurable preference signals. Experiments across four LLMs and seven typologically diverse languages show that MACRO improves *validity* by 12.55% on average over the chain-of-thought baseline without degrading *minimality*, while avoiding the severe *minimality* violations of the translation-based baseline. Compared to supervised fine-tuning, MACRO achieves superior performance on both metrics, confirming that explicit preference optimization is essential for balancing this trade-off. Further analyses reveal that MACRO increases cross-lingual perturbation alignment and mitigates common generation errors. Our results highlight preference optimization as a promising direction for enhancing multilingual model explanations.

1 Introduction

The necessity of multilingual explanations for interpreting black-box LLM behavior has recently been recognized (Resck et al., 2025; Kastrati et al., 2025; Zhao et al., 2025), driven by both empirical evidence of systematic behavioral differences between English and non-English contexts (Lai et al., 2023) and equitable access requirements for non-English-speaking populations (Steigerwald et al., 2022). Among various self-explanations, SCEs offer a causally grounded probe of model behavior: given an input, an LLM generates a minimally modified version (*minimality*) that changes its own prediction (*validity*) (Ross et al., 2021; Bhan et al., 2023; Zhao et al., 2024). However, extending SCEs to multilingual settings poses two key difficulties: ① existing methods largely fail to generate valid SCEs in non-dominant languages (Wang et al., 2026a); ② a trade-off between *validity* and *minimality* may persist (Mayne et al., 2025).[2] We argue that preference optimization (PO) offers a natural solution to these challenges, as it can directly align model outputs toward *validity* while penalizing excessive edits to preserve *minimality*. Despite this potential, applying PO to SCE generation is non-trivial, primarily due to the absence of high-quality multilingual preference data and the difficulty of mathematically formulating the trade-off into measurable preference signals.

To this end, we introduce MACRO, an automated preference alignment framework that leverages DPO (Rafailov et al., 2023) to optimize multilingual SCE generation (Figure 1), turning the need for high *validity* and *minimality* directly into scoring components (§3.2) that are subsequently used for preference pair construction. Our experiments across four LLMs and seven typologically diverse languages demonstrate that MACRO effectively mitigates the *validity–minimality* trade-off that persists in prior approaches (Mayne et al., 2025). Our main contributions are as follows:

First, we show that MACRO consistently improves *validity* over three representative baselines: DG-CF, TB-CF (Wang et al., 2026a), and supervised fine-tuning (SFT). This improvement is

[1] https://anonymous.4open.science/r/MACRO-SCE/

[2] SCEs could be more valid and effective when more modifications are applied. However, deviating too far from the original inputs makes SCEs unreliable as model explanations.

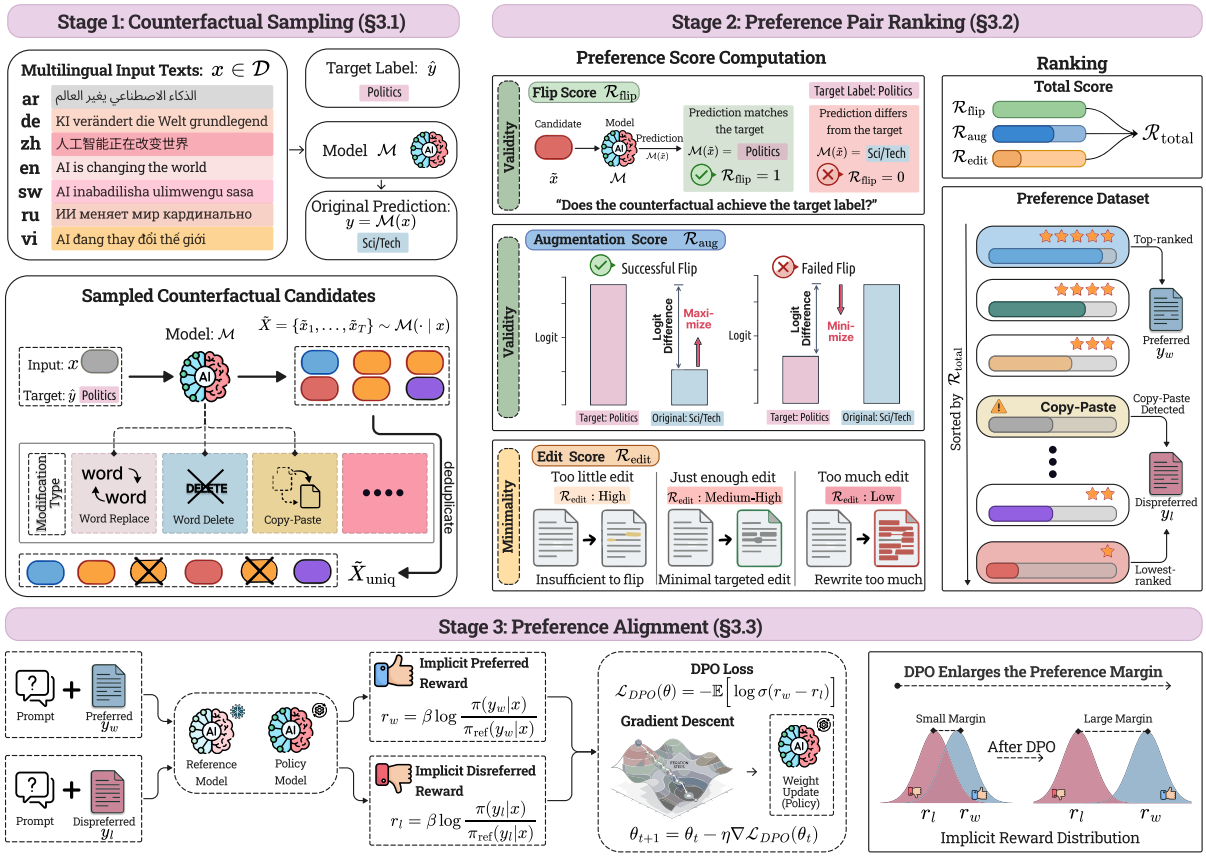


Figure 1: Overview of our three-stage framework (MACRO). **Stage 1** samples counterfactual candidates via word-level perturbations across multilingual inputs. **Stage 2** ranks candidates using $\mathcal{R}_{\text{flip}}$, $\mathcal{R}_{\text{aug}}$, and $\mathcal{R}_{\text{edit}}$ to construct preference pairs. **Stage 3** applies DPO to align the model toward generating minimal, effective counterfactuals.

achieved without degrading *minimality*, marking a pronounced distinction from the translation-based approach TB-CF and SFT baselines.

Second, we confirm that each component in MACRO plays a distinct and complementary role in balancing *validity* and *minimality* by conducting systematic ablation studies on three preference scoring components: flip score ($\mathcal{R}_{\text{flip}}$), augmentation score ($\mathcal{R}_{\text{aug}}$), and edit score ($\mathcal{R}_{\text{edit}}$).

Third, we find that MACRO improves the alignment of cross-lingual input perturbation strategies across all language pairs, by examining whether SCEs across different languages rely on similar types of modifications to change the semantics of the original input.

Lastly, we show that two common issues, *copy-paste* and *language confusion*, can be largely mitigated by MACRO. This effect is especially pronounced for lower-resource languages.

2 Background and Related Work

Multilingual Counterfactuals. Liu et al. (2021) enhance the machine translation performance via multilingual counterfactual data augmentation. Barriere and Cifuentes (2024a,b) evaluate nationality bias in diverse languages using multilingual counterfactual examples. Saha Roy et al. (2025) utilize multilingual counterfactuals to perform answer attribution in multilingual retrieval-augmented generation systems. Nonetheless, the counterfactuals used in the aforementioned work differ from the counterfactual explanations investigated in this paper: while the former flip ground-truth labels, the latter aim to alter the model's predictions. Wang et al. (2026a) introduce the first multilingual counterfactual explanation benchmark and propose two CoT-based approaches (Wei et al., 2022), DG-CF and TB-CF; however, both methods exhibit limited effectiveness, particularly for low-resource languages. Consequently, we explore the potential of PO to enhance multilingual SCE generation and largely mitigate the *validity-minimality* trade-off.

Preference Optimization for Explainability. Preference optimization methods have recently been employed to guide LLMs toward generat-

ing superior outputs compared to inferior alternatives across various tasks (Rafailov et al., 2023; Shao et al., 2024; Ethayarajh et al., 2024; Liu et al., 2026). Among these approaches, Direct Preference Optimization (Rafailov et al., 2023) has gained widespread adoption for LLM alignment, owing to its simplicity. DPO aligns model outputs with human preferences by utilizing both preferred responses (y_w) and dispreferred responses (y_l), with the objective function formulated as:

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{DPO}}}\Big[\log\sigma\big(\beta\log\frac{\pi(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big)\Big] \quad (1)$$

where $\sigma(\cdot)$ represents the sigmod function and β denotes the weighting factor of the Kullback-Leibler divergence (Kullback and Leibler, 1951), which constrains the policy π_θ to remain close to the reference policy π_{ref}.[3] While DPO has proven effective in enhancing natural language explanation quality (Villa-Arenas et al., 2024; Trivedi et al., 2024; Bansal et al., 2026), its use in (multilingual) counterfactual generation is still unexplored. Counterfactual generation demands objective measures of both *minimality* and *validity*, which may involve a trade-off between one another (Mayne et al., 2025). This makes it considerably more challenging compared to other explainability methods.

3 Methodology: MACRO

Given a model $\mathcal{M}$ and an input text $x^{(\ell)} \in \mathcal{D}$ predicted as $y^{(\ell)} = \mathcal{M}(x^{(\ell)})$, where $\mathcal{D}$ is a multilingual dataset with languages $\ell \in \mathcal{L}$ and parallel data, our objective is to construct a counterfactual $\tilde{x}^{(\ell)}$ that causes the model to output the target label $\hat{y}^{(\ell)}$, while maintaining similarity to $x^{(\ell)}$. To address this, we propose MACRO (Figure 1), which applies DPO to multilingual SCE generation, treating *validity* as the preferred outcome while penalizing excessive edits to preserve *minimality*. MACRO operates in three stages: counterfactual sampling, preference pair ranking, and preference alignment.

3.1 Counterfactual Sampling

In the first step, for each data point $x^{(\ell)} \in \mathcal{D}$ across all investigated languages $\ell \in \mathcal{L}$, the model $\mathcal{M}$ samples a set of T diverse counterfactuals, denoted $\tilde{X}^{(\ell)} = \{\tilde{x}_1^{(\ell)}, ..., \tilde{x}_T^{(\ell)}\}$, by applying the CoT-based approach DG-CF (Wang et al., 2026a) to perturb the original input $x^{(\ell)}$ stepwise (for more details, see Section 4.3; the prompt is shown in Figure 9):

$$\tilde{X}^{(\ell)} \sim \mathcal{M}(\cdot \mid x^{(\ell)}) \quad (2)$$

3.2 Preference Pair Ranking

After sampling, we deduplicate the sampled counterfactuals $\tilde{X}^{(\ell)}$, as we find that LLMs often produce identical outputs under stochastic decoding, yielding $\tilde{X}^{(\ell)}_{\text{uniq}}$. We then compute the scores for each candidate and rank them accordingly. The score comprises three components: $\mathcal{R}_{\text{flip}}$, $\mathcal{R}_{\text{aug}}$, and $\mathcal{R}_{\text{edit}}$, covering the *validity* and *minimality* perspectives. $\mathcal{R}_{\text{flip}}$ provides the primary binary signal driving label change (§3.2.1), $\mathcal{R}_{\text{aug}}$ refines this signal with continuous logit-margin feedback to consolidate and strengthen flips (§3.2.2), and $\mathcal{R}_{\text{edit}}$ constrains the edit path to ensure that validity is achieved through minimal modifications (§3.2.3).

3.2.1 Flip Score ($\mathcal{R}_{\text{flip}}$)

One of the core objectives is to ensure that the counterfactual $\tilde{x}$ effectively shifts the model's output from the original prediction y to the desired target label $\hat{y}$ (*validity*). We therefore design $\mathcal{R}_{\text{flip}}$ as a sparse binary score:

$$\mathcal{R}_{\text{flip}}(\tilde{x}, \hat{y}) = \mathbb{I}[\mathcal{M}(\tilde{x}) = \hat{y}] \quad (3)$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 when the condition is satisfied and 0 otherwise.

3.2.2 Augmentation Score ($\mathcal{R}_{\text{aug}}$)

For *validity*, we further introduce an augmentation score to encourage the generated counterfactual either (i) to separate the target label $\hat{y}$ from the original label y by a large margin when the label change succeeds, or (ii) to reduce the model's confidence in the original label when the label change fails, making the candidate closer to a successful label change. The logit vector $\mathbf{z}(x) \in \mathbb{R}^{|V|}$ is the output of the model $\mathcal{M}$ for an input text x, where $|V|$ represents the vocabulary size and $z_c(x)$ denotes the logit of class c.

$$\mathcal{R}_{\text{aug}}(x, \tilde{x}; \hat{y}, y) = \begin{cases} \sigma\big(z_{\hat{y}}(\tilde{x}) - z_y(\tilde{x})\big), \\ \quad \text{if } \mathcal{M}(\tilde{x}) = \hat{y}, \\ \sigma(z_y(x) - z_y(\tilde{x})), \\ \quad \text{otherwise.} \end{cases} \quad (4)$$

3.2.3 Edit Score ($\mathcal{R}_{\text{edit}}$)

The other objective beyond validity is to minimize modifications to the original input x. Following Dehghanighobadi et al. (2025); Wang et al. (2025); Bhattacharjee et al. (2024), we employ a normalized Levenshtein distance d, which captures all

[3] More information about DPO is detailed in Appendix A.

applied edits, converted to a similarity score[4]:

$$\mathcal{R}_{\text{edit}}(x, \tilde{x}) = 1 - \frac{d(x, \tilde{x})}{|x|} \quad (5)$$

This will penalize $\tilde{x}$ with extensive modifications to the input x, which may even result in a negative edit score.

3.2.4 Total Score ($\mathcal{R}_{\text{total}}$)

We compute a scalar total score by a weighted sum over available components:

$$\mathcal{R}_{\text{total}} = \sum_k w_k \mathcal{R}_k, \quad w_k > 0 \quad (6)$$

where k indexes applicable scores with weights w_k and scores $\mathcal{R}_k \in \{\mathcal{R}_{\text{flip}}, \mathcal{R}_{\text{aug}}, \mathcal{R}_{\text{edit}}\}$ (App. B.4).

3.3 Preference Alignment

We construct preference pairs only for examples where at least one sampled candidate changes the model prediction; otherwise, the example is discarded. The preferred candidate $y_w^{(\ell)}$ and the dispreferred candidate $y_l^{(\ell)}$ for input $x^{(\ell)}$ are determined by $\mathcal{R}_{\text{total}}$:

$$y_w^{(\ell)} = \underset{\tilde{x}_j^{(\ell)} \in \tilde{X}_{\text{uniq}}^{(\ell)}}{\arg\max} \mathcal{R}_{\text{total}}(\tilde{x}_j^{(\ell)}) \quad (7)$$

$$y_l^{(\ell)} = \underset{\tilde{x}_j^{(\ell)} \in \tilde{X}_{\text{uniq}}^{(\ell)}}{\arg\min} \mathcal{R}_{\text{total}}(\tilde{x}_j^{(\ell)}) \quad (8)$$

Additionally, if any candidate $\tilde{x}^{(\ell)}$ exactly matches the original input $x^{(\ell)}$, we explicitly set $y_l^{(\ell)} = x^{(\ell)}$ while keeping $y_w^{(\ell)}$ as defined in Eq. (7), thereby discouraging the model from copying the input and facilitating meaningful edits. This yields the DPO training dataset for a given language ℓ, $\mathcal{D}_{\text{DPO}}^{(\ell)} = \{(x^{(\ell)}, y_w^{(\ell)}, y_l^{(\ell)})_i\}_{i=1}^N$, of size N. Leveraging the multilingual DPO dataset $\mathcal{D}_{\text{DPO}} = \cup_{\ell \in \mathcal{L}} \mathcal{D}_{\text{DPO}}^{(\ell)}$, we optimize the model $\mathcal{M}$ with respect to $\mathcal{R}_{\text{total}}$, employing the loss function defined in Eq. (1).[5]

4 Experimental Setup

4.1 Datasets

Consistent with Wang et al. (2026a) and owing to the inherent definition of counterfactual explanation (§3), we focus on two classification tasks and employ the following two large-scale multilingual datasets with sufficient language coverage accordingly for model preference alignment.[6]

SIB200 (Adelani et al., 2024) is a dataset for *topic classification* in 205 languages, designed to evaluate natural language understanding across diverse languages. Sentences are categorized into 7 classes: *Science/Technology*, *Travel*, *Politics*, *Sports*, *Health*, *Entertainment*, and *Geography*.

TAXI1500 (Ma et al., 2025) is a massively multilingual *sentence classification* dataset covering 1504 languages. It is built from parallel Bible translations by annotating English New Testament verses with 6 topics and projecting labels to other languages via verse-level alignment. The 6 classes are *Recommendation*, *Faith*, *Description*, *Sin*, *Grace*, and *Violence*.

Language Selection. We identify seven overlapping languages between SIB200 and TAXI1500: *English*, *German*, *Chinese*, *Arabic*, *Swahili*, *Vietnamese*, and *Russian*. This selection exemplifies typological diversity across language families, encompassing both high- and low-resource languages across diverse scripts.

4.2 Models

In our experiments, we evaluate MACRO on four state-of-the-art open-source multilingual LLMs ranging from 4B to 12B parameters across two model families: `Qwen3-{4B,8B}` (Yang et al., 2025) and `Gemma3-{4B,12B}` (Kamath et al., 2025), all of which support the languages investigated in this paper (§4.1).[7]

4.3 Baselines

Wang et al. (2026a) propose two baselines for multilingual SCE generation: DG-CF and TB-CF. ① DG-CF is based on one-shot CoT prompting and generates SCEs directly in the target language by (1) identifying the most important words and (2) finding appropriate replacement words and substituting them (Appendix Figure 9). As Wang et al. (2026a) report that SCEs in English achieve the highest validity, ② TB-CF first generates English SCEs using DG-CF, and then translates them into the target language using the same LLM. ③ For a fair comparison, we further apply SFT with $\mathcal{D}_{\text{SFT}} = \cup_{\ell \in \mathcal{L}} \mathcal{D}_{\text{SFT}}^{(\ell)}$, where $\mathcal{D}_{\text{SFT}}^{(\ell)} = \{(x^{(\ell)}, y_w^{(\ell)})_i\}_{i=1}^N$, i.e, using only preferred answers (§3.3), *without* dispreferred answers. MACRO uses richer (internal) signals from the base model (§3.2)

[4] Although Levenshtein distances are not mathematically comparable across languages due to differing script systems, they remain comparable within a given language.

[5] Details about DPO training are presented in Appendix B.

[6] Dataset information is detailed in Appendix C.

[7] Model information is reported in Appendix D.

to explicitly optimize *validity* while employing *minimality* regularization, whereas the other baselines either do not use these signals or do not exploit them to the same extent.

4.4 Automatic Evaluation Metrics

Following prior work on counterfactual explanations (Nguyen et al., 2024; Wang et al., 2026b; Ross et al., 2021; Dehghanighobadi et al., 2025), we conduct automatic evaluation to assess the generated multilingual SCEs along three dimensions: (i) *validity*; (ii) *fluency*; and (iii) *minimality*.

Soft Label Flipping Rate (SLFR) measures validity, defined as the fraction of SCEs that successfully change the model prediction.[8] Due to the self-explanation property, the same model $\mathcal{M}$ is used both for generating SCE $\tilde{x}$ and for evaluating whether the predicted label is shifted. Formally, over N evaluation instances, SLFR is computed as:

$$\text{SLFR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big[\mathcal{M}(\tilde{x}_i) \neq \mathcal{M}(x_i)\big]$$

Fluency We assess fluency using perplexity score (PPL) assigned by `mGPT`[9] (Shliazhko et al., 2024), an open-source multilingual autoregressive language model. `mGPT` is selected because it excels at modeling text distributions and covers the support of all investigated languages. For a token sequence $\tilde{x} = (w_1, \ldots, w_T)$, PPL is defined as follows:

$$\text{PPL}(\tilde{x}) = \exp\left\{-\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(w_t \mid w_{<t})\right\}$$

where p_θ denotes the next-token probability.

Semantic Similarity (SS) Minimality can be assessed by quantifying the semantic similarity between the original input x and its counterfactual $\tilde{x}$ with multilingual sentence embeddings, thereby enabling mathematically grounded cross-lingual comparability across the languages studied in this work. We compute SS as the average cosine similarity between x and $\tilde{x}$ over N instances:

$$\text{SS} = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{e}(x_i)^\top \mathbf{e}(\tilde{x}_i)}{\|\mathbf{e}(x_i)\|_2 \, \|\mathbf{e}(\tilde{x}_i)\|_2}$$

where $\mathbf{e}(\cdot)$ denotes the sentence embedding.[10]

[8] We emphasize that SLFR checks whether $\tilde{x}$ has a different prediction $\mathcal{M}(\tilde{x})$ than y and is thus different from the flip score (§3.2.1), which checks whether $\tilde{x}$ reaches the target label $\hat{y}$, where $\hat{y}$ is not necessarily equal to $\mathcal{M}(\tilde{x})$.

[9] https://huggingface.co/ai-forever/mGPT

[10] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

4.5 Cross-lingual Edit Similarity

Cross-lingual Edit Similarity (CES) measures the degree to which semantic modifications in counterfactuals $\tilde{x}^{(\ell)}$ are consistently applied across language pairs (ℓ_1, ℓ_2) relative to the original input (Wang et al., 2026a):

$$\text{CES}(\ell_1, \ell_2) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{e}(\tilde{x}_i^{(\ell_1)})\top \mathbf{e}(\tilde{x}_i^{(\ell_2)})}{\|\mathbf{e}(\tilde{x}_i^{(\ell_1)})\|_2 \, \|\mathbf{e}(\tilde{x}_i^{(\ell_2)})\|_2}$$

Higher values indicate that edits are transferred more uniformly between the two languages. We adopt CES to examine how perturbation consistency in multilingual SCE generation evolves through MACRO.

5 Results

5.1 Multilingual Counterfactual Quality

MACRO generally outperforms DG-CF and SFT. Table 1 shows that MACRO achieves average relative improvements of 12.55% in *validity*, 18.63% in *fluency*, and 3.13% in *minimality* over DG-CF. This highlights MACRO's ability to mitigate the *validity-minimality* trade-off effectively (Figure 2). Notably, improvements are more substantial on TAXI1500, likely because its parallel Bible verses construction provides a more predictable syntactic and semantic structure that facilitates counterfactual optimization. While larger models generally outperform smaller counterparts within families, performance differences between high- and relatively low-resource languages (ar, vi, sw) remain surprisingly marginal. Moreover, no single language invariably yields optimal SCEs across all metrics. Regarding cross-lingual stability, `Gemma3-12B` exhibits the lowest *validity* variation across seven languages, whereas `Qwen3-8B` shows the poorest cross-lingual stability, driven primarily by its suboptimal performance on Swahili. Additionally, MACRO outperforms SFT on both *validity* (+1.3%) and *minimality* (+3.5%) on average. While SFT occasionally surpasses MACRO on *validity* on SIB200, its consistent underperformance on *minimality* (frequently even worse than the DG-CF baseline) makes it less suitable for faithful SCE generation. This confirms that the trade-off cannot be easily resolved through standard supervised learning alone and preference optimization proves essential for achieving Pareto-optimal SCE quality.[11]

[11] Appendix I further demonstrates that, compared to SFT, MACRO better preserves or even enhances the model's general and multilingual capabilities while simultaneously enhancing cross-lingual generalizability in terms of SCE validity.

Dataset		SIB200									TAXI1500								
Model	Lang-uage	Validity/SLFR (↑)			Fluency/PPL (↓)			Minimality/SS (↑)			Validity/SLFR (↑)			Fluency/PPL (↓)			Minimality/SS (↑)		
		DG-CF	SFT	MACRO	DG-CF	SFT	MACRO	DG-CF	SFT	MACRO	DG-CF	SFT	MACRO	DG-CF	SFT	MACRO	DG-CF	SFT	MACRO
Gemma3-4B	en	0.63	0.70	0.64	48.8	31.4	31.7	0.61	0.55	0.65	0.53	0.56	0.60	36.2	23.4	23.5	0.51	0.59	0.59
	de	0.59	0.65	0.64	28.5	24.0	24.4	0.75	0.71	0.78	0.56	0.59	0.66	33.9	25.2	23.0	0.64	0.69	0.73
	zh	0.66	0.76	0.80	35.1	28.5	30.1	0.69	0.65	0.66	0.52	0.61	0.60	39.5	30.3	32.4	0.61	0.65	0.67
	ar	0.61	0.74	0.66	27.7	24.6	26.3	0.66	0.60	0.65	0.52	0.54	0.57	41.9	32.8	30.5	0.63	0.60	0.66
	vi	0.59	0.65	0.65	18.9	16.5	18.7	0.71	0.67	0.71	0.50	0.51	0.51	18.5	17.5	16.3	0.61	0.64	0.66
	sw	0.68	0.72	0.74	18.7	18.2	17.6	0.68	0.69	0.67	0.57	0.57	0.64	25.6	20.6	19.9	0.67	0.75	0.76
	ru	0.61	0.72	0.64	20.3	19.3	19.1	0.74	0.70	0.74	0.53	0.54	0.60	21.5	17.9	19.0	0.66	0.65	0.68
	avg	0.62	0.71	0.68	28.3	23.2	24.0	0.69	0.65	0.69	0.53	0.56	0.60	31.0	24.0	23.5	0.62	0.65	0.68
Qwen3-4B	en	0.54	0.68	0.66	53.4	32.8	34.8	0.79	0.68	0.75	0.62	0.63	0.66	32.6	25.1	26.1	0.61	0.60	0.68
	de	0.62	0.65	0.66	44.4	26.9	27.9	0.71	0.71	0.76	0.49	0.54	0.53	40.2	26.8	27.0	0.66	0.67	0.70
	zh	0.81	0.78	0.72	47.7	31.7	33.5	0.60	0.61	0.73	0.59	0.66	0.60	45.8	36.9	39.1	0.67	0.65	0.71
	ar	0.64	0.72	0.66	43.3	29.8	32.2	0.70	0.59	0.70	0.68	0.69	0.72	86.2	43.6	48.6	0.58	0.53	0.60
	vi	0.71	0.73	0.71	25.0	18.9	21.7	0.69	0.65	0.71	0.65	0.60	0.66	25.7	21.5	20.7	0.66	0.67	0.67
	sw	0.49	0.47	0.41	31.2	21.6	21.2	0.73	0.78	0.80	0.46	0.48	0.51	25.6	18.8	17.7	0.84	0.84	0.83
	ru	0.45	0.69	0.62	26.8	21.2	22.0	0.84	0.72	0.77	0.53	0.57	0.64	27.7	23.1	23.2	0.68	0.67	0.71
	avg	0.61	0.67	0.63	38.8	26.1	27.6	0.72	0.68	0.75	0.57	0.60	0.62	40.5	28.0	28.9	0.67	0.66	0.70
Qwen3-8B	en	0.59	0.70	0.74	53.2	33.0	37.9	0.73	0.62	0.66	0.56	0.70	0.72	31.0	20.7	25.1	0.68	0.61	0.69
	de	0.65	0.75	0.73	32.1	24.6	27.8	0.69	0.65	0.71	0.50	0.62	0.71	25.9	21.4	22.9	0.62	0.65	0.66
	zh	0.62	0.72	0.70	38.6	30.0	34.9	0.68	0.67	0.70	0.63	0.63	0.66	38.9	30.2	33.6	0.66	0.63	0.71
	ar	0.58	0.77	0.76	34.9	26.8	30.5	0.74	0.60	0.65	0.66	0.84	0.76	50.3	37.0	37.8	0.64	0.59	0.63
	vi	0.68	0.73	0.72	22.7	18.3	23.1	0.69	0.65	0.71	0.58	0.70	0.78	18.8	16.0	18.0	0.61	0.60	0.64
	sw	0.53	0.56	0.57	22.9	18.5	19.7	0.84	0.78	0.79	0.11	0.32	0.27	19.0	17.4	16.8	0.83	0.79	0.83
	ru	0.64	0.72	0.66	22.1	19.6	21.5	0.70	0.64	0.71	0.45	0.56	0.61	18.7	15.6	20.2	0.68	0.66	0.73
	avg	0.61	0.71	0.70	32.4	24.4	27.9	0.72	0.66	0.70	0.50	0.62	0.64	28.9	22.6	24.9	0.67	0.65	0.70
Gemma3-12B	en	0.83	0.90	0.84	53.9	34.3	38.3	0.53	0.49	0.51	0.83	0.81	0.86	41.5	26.7	27.9	0.44	0.45	0.49
	de	0.77	0.84	0.85	28.1	24.1	27.0	0.57	0.58	0.58	0.79	0.77	0.81	26.5	21.8	21.7	0.54	0.59	0.60
	zh	0.84	0.81	0.92	31.6	29.7	32.6	0.55	0.57	0.51	0.68	0.60	0.76	27.9	26.0	28.9	0.54	0.56	0.59
	ar	0.87	0.87	0.92	25.5	25.1	26.6	0.52	0.55	0.49	0.78	0.79	0.79	37.3	28.1	36.3	0.54	0.55	0.59
	vi	0.79	0.79	0.80	18.5	17.7	21.9	0.56	0.59	0.57	0.74	0.69	0.80	16.2	15.1	17.1	0.52	0.55	0.56
	sw	0.81	0.82	0.83	14.9	14.7	15.9	0.65	0.68	0.71	0.69	0.62	0.73	17.5	16.7	15.9	0.68	0.73	0.75
	ru	0.78	0.79	0.80	21.0	19.6	22.1	0.59	0.57	0.61	0.75	0.69	0.81	20.5	16.3	18.5	0.60	0.61	0.61
	avg	0.81	0.83	0.85	27.6	23.6	26.3	0.57	0.58	0.57	0.75	0.71	0.79	26.8	21.5	23.8	0.55	0.58	0.60

Table 1: Automatic evaluation results on SIB200 and TAXI1500. We compare DG-CF, SFT, and MACRO across three metrics: soft label flipping rate (SLFR (↑)), perplexity (PPL (↓)), and semantic similarity (SS (↑)). The **best** and second-best results for each language-metric pair are highlighted accordingly. Results for TB-CF are provided in Appendix Table 14 and discussed in §5.1.

MACRO produces more reliable and stable SCEs than TB-CF. Appendix Table 14 shows that, while TB-CF achieves a higher SLFR compared to DG-CF, SFT, and MACRO, it requires substantially more modifications, as indicated by markedly reduced *minimality*. Specifically, its Edit Score ($\mathcal{R}_{\text{edit}}$; §3.2.3) is frequently below 0, suggesting that the edit length surpasses that of the original input. Manual inspection reveals that these excessive modifications primarily stem from translation artifacts rather than semantically meaningful changes (Tables 8 and 9). *This violates the minimality constraint of counterfactual explanations* (§3). Consequently, despite TB-CF attaining high SLFR (*validity*), it strays noticeably from the original input without pinpointing the nuanced feature changes that account for the explained model's decisions (*minimality*), severely undermining their reliability as SCEs.

Validity-Minimality Trade-off. The trade-off between *validity* and *minimality* observed by Mayne et al. (2025) can be ascribable to the imperfect SCE generation approaches, which are mostly prompting-based without robust regularization to ensure the *minimality* (Cardiel et al., 2026), as also evidenced by our ablation study (§5.2). Given that we incorporate three meticulously designed scoring components aimed at optimizing *validity* and *minimality* (§3.2), MACRO is able to enhance *validity* by 12.55% without sacrificing *minimality* (even enhancing *minimality* by 3.13% on average), which largely mitigates the previously observed trade-off (Mayne et al., 2025) (Figure 2, Appendix J).

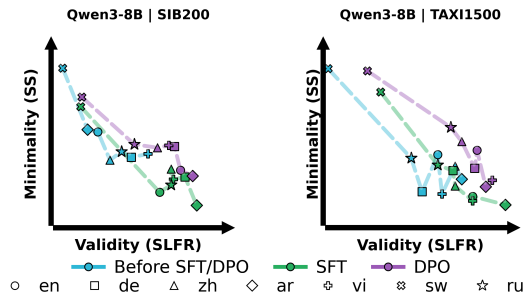


Figure 2: The validity-minimality trade-off across languages with Qwen3-8B on SIB200 and Taxi1500.

English is not necessarily the optimal language for counterfactual generation. Table 1 reveals that English counterfactuals do not consistently achieve the highest validity across all experimental settings. This observation contradicts the findings of Wang et al. (2026a), who report that counterfactuals in high-resource European languages generally exhibit superior quality. We argue that the optimal language is highly dependent on the explained model and its internal decision boundaries for different languages, which vary across models.

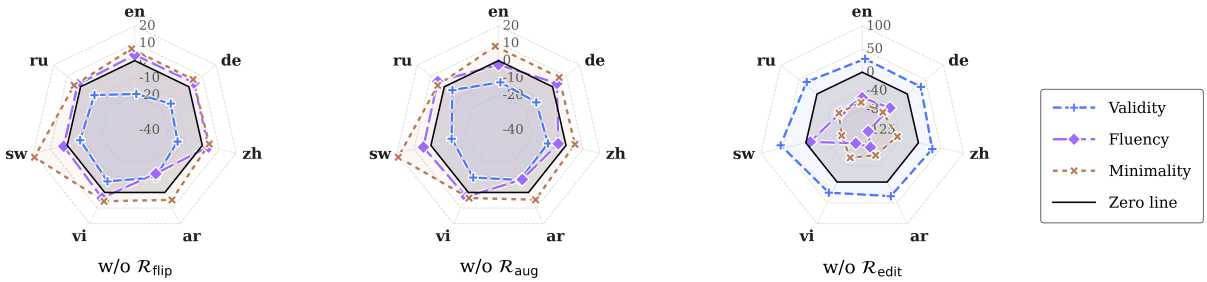


Figure 3: Relative performance change across languages for `Qwen3-8B` under three ablation settings. We evaluate the removal of $\mathcal{R}_{\text{flip}}$, $\mathcal{R}_{\text{aug}}$, and $\mathcal{R}_{\text{edit}}$ across *validity*, *fluency*, and *minimality*. Each radar plot illustrates results for a specific ablation, with scores averaged over SIB200 and TAXI1500. *Positive* values denote a percentage *improvement* in performance, while *negative* values indicate a *degradation* relative to the full MACRO configuration.

5.2 Ablation Study

MACRO employs three scoring components for preference pair ranking (§3.2): $\mathcal{R}_{\text{flip}}, \mathcal{R}_{\text{aug}}, \mathcal{R}_{\text{edit}}$, which collectively aim to optimize for *validity* and *minimality* of counterfactuals. Accordingly, to demonstrate their individual contributions, we conduct an ablation study for each component.[12] Figure 3 and Appendix Table 15 present the ablation study results for `Qwen3-8B`, obtained by removing one scoring component (§3.2) at a time.[13]

Effect of Removing $\mathcal{R}_{\text{flip}}$. Removing $\mathcal{R}_{\text{flip}}$ leads to the most pronounced drop in *validity*, with an average absolute decline of 11.65% compared to the full model (Appendix Table 18). Meanwhile, the removal of $\mathcal{R}_{\text{flip}}$ yields moderate drops in fluency and improvement in *minimality*. This occurs because removing $\mathcal{R}_{\text{flip}}$ allows $\mathcal{R}_{\text{edit}}$ to dominate the ranking, preferentially selecting candidates with fewer modifications regardless of whether they achieve a successful label flip. The result confirms that an explicit label-flip constraint is essential; without it, the model gravitates toward the original input and fails at the primary objective of SCE generation.

Effect of Removing $\mathcal{R}_{\text{aug}}$. *Validity* decreases notably without $\mathcal{R}_{\text{aug}}$, although the degradation is less severe than w/o $\mathcal{R}_{\text{flip}}$ (Appendix Table 19), confirming that $\mathcal{R}_{\text{aug}}$ serves as a complementary objective to $\mathcal{R}_{\text{flip}}$. Mechanistically, $\mathcal{R}_{\text{aug}}$ is designed to facilitate margin maximization (§3.2.2), effectively "nudging" borderline SCEs across the decision boundary. In its absence, these near-miss candidates often fail to achieve a successful label flip. Similar to w/o $\mathcal{R}_{\text{flip}}$, *minimality* is generally enhanced without $\mathcal{R}_{\text{aug}}$.

Effect of Removing $\mathcal{R}_{\text{edit}}$. The removal of $\mathcal{R}_{\text{edit}}$ results in a markedly different failure mode. *Validity* consistently increases across all languages relative to the full model. However, this apparent improvement in *validity* comes at a severe cost to *minimality* (Appendix Table 20), revealing that the model achieves label flips by making excessively large modifications and deviating significantly from the original semantics. These findings demonstrate that, without $\mathcal{R}_{\text{edit}}$, the model degenerates into producing heavily modified outputs that trivially flip labels but fail to satisfy the *minimality* criterion central to counterfactual explanations.

Language-Specific Observations. Appendix Table 21 illustrates that **Swahili** emerges as the *most* sensitive language overall, particularly in its extreme minimality and validity sensitivity. In comparison, **Russian** is the *least* sensitive language, with the lowest validity sensitivity. **English** and **German** display the most balanced sensitivity profiles, with all three dimensions falling within a relatively narrow range, confirming that their label-flip success depends most critically on the collaboration among all three scoring components (Appendix M.3).

5.3 Cross-lingual Edit Similarity

Before applying MACRO. Figures 4 and 15 reveal that, in alignment with the findings of Wang et al. (2026a), closely related languages (German, English, and Russian) achieve moderately higher CES, suggesting that semantic edits in counterfactuals are applied with considerable consistency across these language pairs, whereas Swahili scores notably low across the board. Surprisingly, although Vietnamese, Arabic, and Chinese are not typologically close to English, German, or Russian,

[12] The discussion of the exclusion of multilingual alignment is provided in Appendix F.

[13] Further results with other models appear in Appendix M.2; example SCEs from each ablation condition are shown in Appendix M.1.

Figure 4: Cross-lingual edit similarity score changes (Δ) for SIB200 and TAXI1500 counterfactuals across language pairs before and after applying MACRO.

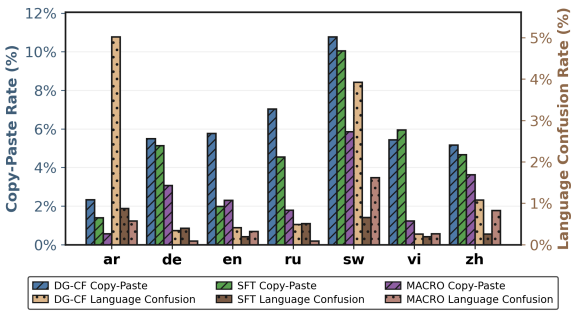


Figure 5: Comparison of *copy-paste* and *language confusion* rates between DG-CF and MACRO.

they achieve relatively high CES scores with them. This indicates that topological similarity between language pairs is not the only influential factor; the resource availability of languages also plays a role.

After applying MACRO. CES improvements are observed across all language pairs on SIB200, with the most substantial gains on high-resource languages. On Taxi1500, the CES of the majority of language pairs is enhanced, but a few pairs involving Swahili exhibit marginal decreases. This may be attributed to the fact that Taxi1500 likely demands finer-grained semantic edits and that MACRO primarily benefits well-represented languages by aligning their counterfactual editing behavior. This optimization may subtly shift the model's internal representations in a direction that marginally disadvantages Swahili pairs.

5.4 Error Analysis

We further report two main failure cases: *copy-paste* and *language confusion*.

Copy-Paste. When asked to generate a counterfactual, the LLM may tend to produce the identical text as the original input, either due to inertia or inability to provide an actionable label flip. Figure 5 shows that MACRO reduces the copy-paste rate by an average of 58.21% across all models and languages compared to DG-CF, since we intentionally designate such SCEs as dispreferred answers (§3.3). This improvement is substantially larger than that of SFT, which achieves only a 19.64% relative reduction over DG-CF. At the language level, the contrast is especially clear in Swahili, where MACRO yields a pronounced reduction in copy-paste rate, while SFT brings only a comparatively modest change.

Language Confusion. We further analyze the language in which SCEs $\tilde{x}$ are generated and assess whether it corresponds to the desired target language.[14] Figure 5 illustrates that MACRO yields clear reductions in language confusion, with an average relative decline of 49.65% across languages, which is particularly pronounced for Arabic and Swahili. Moreover, MACRO noticeably reduces the tendency to generate SCEs in non-targeted languages (most often English, given the models' English-centric properties), thereby improving consistency in targeted language generation.

6 Conclusion

In this work, we proposed MACRO, a preference alignment framework that leverages DPO to optimize multilingual SCE generation. Our approach addresses two persistent challenges in extending SCEs beyond English: the failure of existing methods to produce valid SCEs in non-dominant languages and the inherent trade-off between *validity* and *minimality*. Experiments across four LLMs and seven languages demonstrate that MACRO consistently improves *validity* over DG-CF by an average of 12.55% without compromising *minimality*, and avoids the excessive modifications introduced by TB-CF. Compared to SFT, MACRO achieves superior performance on both metrics, confirming that explicit preference optimization is essential to mitigate this trade-off. Ablation studies further confirm the complementary roles of the three preference scoring components in shaping effective preference signals. In addition, we showed that MACRO promotes greater cross-lingual alignment in input perturbation strategies and substantially mitigates common failure modes such as *copy-paste* and *language confusion* errors. Our work highlights that preference optimization is a promising paradigm for improving the quality of model self-explanations in multilingual settings.

[14] https://github.com/cisnlp/GlotLID

Limitations

We acknowledge several limitations of this work:

Limited Exploration of Language. This paper focuses only on seven selected languages (§4.1). While these languages represent diverse typological properties and resource levels, the effectiveness of MACRO on unexplored languages remains to be established.

Limited Exploration RL Approach. In this paper, we investigate only the potential of DPO to enhance (multilingual) counterfactual explanation generation.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Pulkit Bansal, Raghvendra Kumar, Shakti Singh, Sriparna Saha, and Adam Jatowt. 2026. From fragments to facts: A curriculum-driven dpo approach for generating hindi news veracity explanations. *Preprint*, arXiv:2507.05179.

Valentin Barriere and Sebastian Cifuentes. 2024a. Are text classifiers xenophobic? a country-oriented bias detection method with least confounding variables. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1511–1518, Torino, Italia. ELRA and ICCL.

Valentin Barriere and Sebastian Cifuentes. 2024b. A study of nationality bias in names and perplexity using off-the-shelf affect-related tweet classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 569–579, Miami, Florida, USA. Association for Computational Linguistics.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2023. Tigtec: Token importance guided text counterfactuals. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pages 496–512, Cham. Springer Nature Switzerland.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Zero-shot LLM-guided Counterfactual Generation: A Case Study on NLP Model Evaluation. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1243–1248, Los Alamitos, CA, USA. IEEE Computer Society.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Amaia Cardiel, Eloi Zablocki, Elias Ramzi, and Eric Gaussier. 2026. Driv-ex: Counterfactual explanations for driving llms. *Preprint*, arXiv:2603.00696.

Zahra Dehghanighobadi, Asja Fischer, and Muhammad Bilal Zafar. 2025. Can LLMs explain themselves counterfactually? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7787–7815, Suzhou, China. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Muhamet Kastrati, Ali Shariq Imran, Ehtesham Hashmi, Zenun Kastrati, Sher Muhammad Daudpota, and Marenglen Biba. 2025. Unlocking language barriers: Assessing pre-trained large language models across multilingual tasks and unveiling the black box with explainable artificial intelligence. *Engineering Applications of Artificial Intelligence*, 149:110136.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585, Toronto, Canada. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Qi Liu, Matt Kusner, and Phil Blunsom. 2021. Counterfactual data augmentation for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 187–197, Online. Association for Computational Linguistics.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *Preprint*, arXiv:2601.05242.

Chunlan Ma, Ayyoob Imani, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schuetze. 2025. Taxi1500: A dataset for multilingual text classification in 1500 languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 414–439, Albuquerque, New Mexico. Association for Computational Linguistics.

Harry Mayne, Ryan Othniel Kearns, Yushi Yang, Andrew M. Bean, Eoin D. Delaney, Chris Russell, and Adam Mahdi. 2025. LLMs don't know their own decision boundaries: The unreliability of self-generated counterfactual explanations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24161–24186, Suzhou, China. Association for Computational Linguistics.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024. LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824, Miami, Florida, USA. Association for Computational Linguistics.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking language comes at the cost of accuracy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20279–20296, Suzhou, China. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20465–20497, Suzhou, China. Association for Computational Linguistics.

Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, Online. Association for Computational Linguistics.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2025. Evidence contextualization and counterfactual attribution for conversational qa over heterogeneous data with rag systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 1040–1043, New York, NY, USA. Association for Computing Machinery.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Emma Steigerwald, Valeria Ramírez-Castañeda, Débora Y C Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D Tarvin. 2022. Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future. *BioScience*, 72(10):988–998.

Prapti Trivedi, Aditya Gulati, Oliver Molenschot, Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Keith Stevens, Tanveesh Singh Chaudhery, Jahnavi Jambholkar, James Zou, and Nazneen Rajani. 2024. Self-rationalization improves llm as a fine-grained judge. *Preprint*, arXiv:2410.05495.

Luis Felipe Villa-Arenas, Ata Nizamoglu, Qianli Wang, Sebastian Möller, and Vera Schmitt. 2024. Anchored alignment for self-explanations enhancement. *Preprint*, arXiv:2410.13216.

Qianli Wang, Nils Feldhus, Simon Ostermann, Luis Felipe Villa-Arenas, Sebastian Möller, and Vera Schmitt. 2025. FitCF: A framework for automatic feature importance-guided counterfactual example generation. In *ACL Findings*, pages 1176–1191, Vienna, Austria. Association for Computational Linguistics.

Qianli Wang, Van Bach Nguyen, Yihong Liu, Fedor Splitt, Nils Feldhus, Christin Seifert, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2026a. Parallel universes, parallel languages: A comprehensive study on llm-based multilingual counterfactual example generation. *Preprint*, arXiv:2601.00263.

Yilong Wang, Qianli Wang, and Nils Feldhus. 2026b. iflip: Iterative feedback-driven counterfactual example refinement. *Preprint*, arXiv:2601.01446.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2).

Li Zhao, Jinwei Zhou, Jinde Cao, and Weina Zhu. 2025. Emsa: Explainable multilingual sentiment analysis models providing sentiment analysis across multiple languages. *PLOS ONE*, 20(11):1–32.

A Direct Preference Optimization

Reward Modeling. Given a dataset $\mathcal{D} = \{(x, y_w, y_l)_i\}_{i=1}^{N}$ with N samples, where x denotes the input prompt and y_w and y_l represent the preferred and dispreferred answers, respectively. r_ϕ is the reward model trained with Bradley-Terry model (Bradley and Terry, 1952) by minimizing the following loss, where $\sigma(\cdot)$ is the sigmoid function:

$$\mathcal{L}_{R(r_\phi,\mathcal{D})} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[log\sigma\big(r_\phi(x,y_w) - r_\phi(x,y_l)\big)\Big] \quad (9)$$

Policy Optimization. The policy π_θ is then optimized to maximize expected rewards guided by the reward function r_ϕ:

$$\mathcal{J}(\theta) = \max_{\pi} \mathbb{E}_{\substack{x\sim\mathcal{D} \\ y\sim\pi(y|x)}}\left[r_\phi(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right] \quad (10)$$

where β denotes the weighting factor of the Kullback-Leibler divergence (Kullback and Leibler, 1951), which constrains the policy π_θ to remain close to the reference policy π_{ref}.

Direct Preference Optimization. DPO is built on a key theoretical insight: under the optimal policy that maximizes the constrained reward objective (Rafailov et al., 2023), there exists a closed-form relationship between the reward function and the policy. Specifically, the optimal reward function can be expressed as:

$$r_\theta(x,y) = \beta \ln \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \ln(Z(x)) \quad (11)$$

where $Z(x) = \sum_y \pi_{ref}(y|x) \cdot \exp(\frac{1}{\beta} r_\phi(x,y))$ is the partition function that ensures the relationship holds across all possible completions. The DPO loss function is derived by substituting the reward function Eq. (11) into the reward modeling objective Eq. (9):

$$\mathcal{L}_{DPO}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\big(\beta \log \frac{\pi(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\big)\Big] \quad (12)$$

B Direct Preference Optimization Training

B.1 Training Time

Table 2 summarizes the total training time for the four base models on SIB200 and TAXI1500.

Model	SIB200	TAXI1500
Gemma3-4B	0:32:47	0:23:11
Qwen3-4B	0:32:21	0:23:43
Qwen3-8B	0:43:22	0:32:56
Gemma3-12B	1:12:23	0:52:09

Table 2: Training time on an NVIDIA H100 GPU for the four model, reported separately on SIB200 and TAXI1500. Times are formatted as hh : mm : ss.

Hyperparameter	**Value**
DPO learning rate	1×10^{-6}
DPO β	0.2
DPO epochs	3
Per-device batch size	7
Gradient accumulation	7
Warmup ratio	0.03
LoRA rank	128
LoRA alpha	256
LoRA dropout	0.05
Weight decay	0.0

Table 3: Default hyperparameter settings.

B.2 Training Detail

For SIB200 and Taxi1500, we construct 600 and 800 preference pairs per language, respectively. We fine-tune the base model using DPO on prompt–chosen–rejected triples. To reduce training cost, we adopt LoRA and insert trainable low-rank adapters into both attention and MLP projection layers. For multilingual training, we employ a language-level round-robin batch sampler to maintain balanced updates across languages. Table 3 summarizes the main hyperparameters.

B.3 Total Score Changes by Direct Preference Optimization

Figure 6 displays the changes in total score distribution after applying MACRO. MACRO can increase overall scores and reduce the variance of the scores while modestly increasing the mean scores after post-training by approximately 9%, indicating that the model becomes more consistent by elevating lower-performance cases without necessarily improving peak performance.

B.4 Scoring Weight Configuration

To compute the total preference score in §3.2.4, we assign weights to the three scoring components, i.e., $\mathcal{R}_{\text{flip}}$, $\mathcal{R}_{\text{aug}}$, and $\mathcal{R}_{\text{edit}}$. Table 4 summarizes the scoring weights used for each model on SIB200 and TAXI1500.

Dataset	Model	w_{flip}	w_{aug}	w_{edit}
SIB200	Qwen3-4B	1.0	0.6	1.0
SIB200	Qwen3-8B	1.0	2.0	1.0
SIB200	Gemma3-4B	1.0	1.35	1.0
SIB200	Gemma3-12B	1.0	1.2	1.0
TAXI1500	Qwen3-4B	1.0	1.0	1.0
TAXI1500	Qwen3-8B	1.0	1.0	1.0
TAXI1500	Gemma3-4B	1.0	1.0	1.0
TAXI1500	Gemma3-12B	1.0	0.5	1.0

Table 4: Scoring weights used to compute the total preference score for each model and dataset.

C Dataset Information

C.1 Dataset Distribution

Figure 7 reports the label distributions of SIB200 and TAXI1500, where percentages are computed over all splits combined.

C.2 Dataset Example

Figures 10a and 10b report representative examples from SIB200 and TAXI1500, respectively.

D Model

D.1 Model Information

Table 5 shows information about the models we employ, including model parameter sizes, citations, and their corresponding Hugging Face link.

D.2 Inference Time

Table 6 reports the inference time of MACRO using Qwen3-8B on SIB200 and TAXI1500, broken down by language.

E Prompt

E.1 Prediction Prompt

Figure 8 reports the prediction prompts used for SIB200 and TAXI1500.

E.2 Counterfactual Generation Prompt

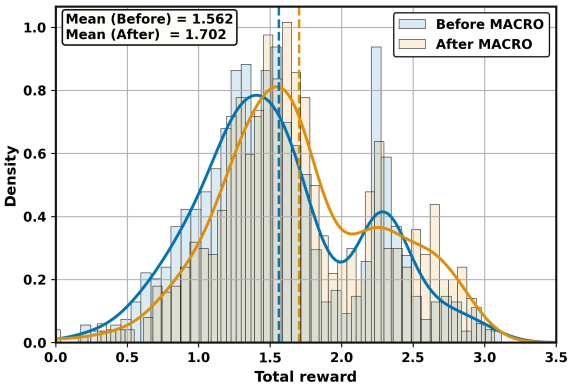
Figure 9 reports counterfactual generation prompts used in our experiments. For a fair comparison, DG-CF and MACRO use the same prompt template.

F Why Not Multilingual Alignment?

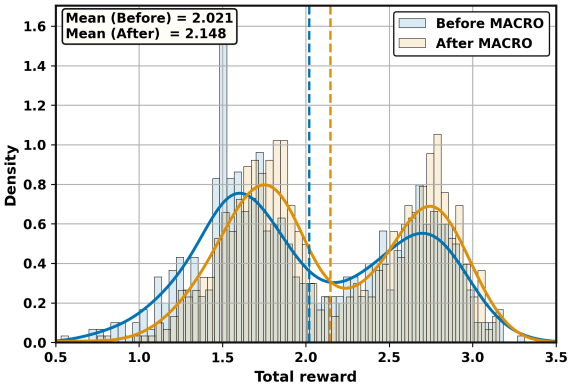
While one might intuitively consider apply multilingual alignment to enhance performance in non-dominant languages, it is challenging to identify a target dominant language with which other languages align, since English is not necessarily the optimal languages (§5.1) unlike in other tasks, e.g.,

Name	Citation	Size	Link
Qwen3	Yang et al. (2025)	4B	https://huggingface.co/Qwen/Qwen3-4B
Qwen3	Yang et al. (2025)	8B	https://huggingface.co/Qwen/Qwen3-8B
Gemma3	Kamath et al. (2025)	4B	https://huggingface.co/google/gemma-3-4b-it
Gemma3	Kamath et al. (2025)	12B	https://huggingface.co/google/gemma-3-12b-it

Table 5: Four open-source LLMs are used for counterfactual generation.



(a) TAXI1500



(b) SIB200

Figure 6: Total score distributions before and after applying MACRO using Qwen3-4B.

Language	SIB200	TAXI1500
en	00:06:50	00:03:09
de	00:09:35	00:04:07
zh	00:07:11	00:03:22
ar	00:10:12	00:03:43
vi	00:09:44	00:04:08
sw	00:10:30	00:04:35
ru	00:11:47	00:04:47

Table 6: Inference time of MACRO using Qwen3-8B on SIB200 and TAXI1500, reported separately by language. Times are formatted as hh:mm:ss.

reasoning (Qi et al., 2025) or translation (Etxaniz et al., 2024). Nevertheless, we follow MAPO (She et al., 2024) to perform cross-lingual alignment using the conditional generation probability from a translation model. We conduct a small-scale ablation study by *including* an additional reward for multilingual alignment and check how it affects the overall counterfactual quality.

F.1 Alignment Score ($\mathcal{R}_{\text{align}}$)

To address the performance gap between counterfactuals in dominant and non-dominant languages (Wang et al., 2026a), we optimize the latter relative to the former. We designate English as the dominant language throughout, given its prevalence in both counterfactual generation and translation. To ensure semantic consistency in multilingual settings, we measure cross-lingual alignment using the conditional generation probability from an off-the-shelf multilingual translation model NLLB-200[15] (Koishekenov et al., 2023), following MAPO (She et al., 2024). Concretely, let $\tilde{x}_{\text{en}}$ denote the English counterfactual anchor and $\tilde{x}_{\text{non}}$ the generated non-English counterfactual. We compute the token-level log-likelihood of generating $\tilde{x}_{\text{en}}$ conditioned on $\tilde{x}_{\text{non}}$ via teacher forcing:

$$\mathcal{L}_{\text{CE}}(\tilde{x}_{\text{en}}, \tilde{x}_{\text{non}}) = -\frac{1}{T}\sum_{t=1}^{T} \log P_{\text{NLLB}}(\tilde{x}_{\text{en}}^{t} \mid \tilde{x}_{\text{en}}^{<t}, \tilde{x}_{\text{non}}) \quad (13)$$

We define the alignment reward $\mathcal{R}_{\text{align}}$ as the NLLB model probability:

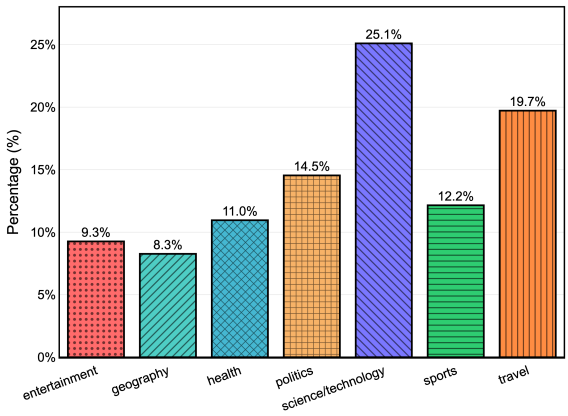
$$\mathcal{R}_{\text{align}}(\tilde{x}_{\text{en}}, \tilde{x}_{\text{non}}) = \sigma\big(-\mathcal{L}_{\text{CE}}(\tilde{x}_{\text{en}}, \tilde{x}_{\text{non}})\big) \quad (14)$$

Higher $\mathcal{R}_{\text{align}}$ indicates greater probability of translating the non-English counterfactual $\tilde{x}_{\text{non}}$ back to the English anchor $\tilde{x}_{\text{en}}$, thus reflecting stronger cross-lingual semantic alignment. Subsequently, we add $\mathcal{R}_{\text{align}}$ to the total reward $\mathcal{R}_{\text{total}}$ (§3.2.4).

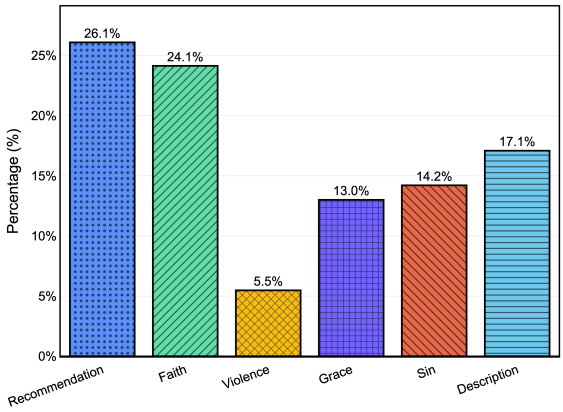
F.2 Results and Discussion

Table 7 shows that although counterfactual quality in terms of *validity* and *fluency* in some languages

[15] https://huggingface.co/facebook/nllb-200-distilled-1.3B



(a) SIB200 label distribution.



(b) TAXI1500 label distribution.

Figure 7: Label distributions of the two evaluation datasets. Subfigure (a) shows SIB200 and subfigure (b) shows TAXI1500.

is enhanced (Swahili and Russian on SIB200), the alignment generally harms the other languages. In particular, *minimality* is noticeably degraded after multilingual alignment. Although one could consistently use the best-performing language as the dominant one, prior knowledge is needed to determine the optimal language on a specific dimension by evaluating DG-CFs, which is *task- and model-specific*. Furthermore, as we reported in §5.1, no single language outperforms others across all evaluation dimensions. These factors make multilingual alignment largely **infeasible** and **ineffective**. Therefore, we discard the use of multilingual alignment.

G Counterfactual Examples

We present multilingual examples from TAXI1500 in Table 8 and from SIB200 in Table 9. These examples illustrate the original inputs and counterfactuals produced by DG-CF, TB-CF, and MACRO.

H Additional Counterfactual Evaluation Metrics

Table 12 reports the additional automatic evaluation metrics of the experiments in Table 1, namely hard label flipping rate and textual similarity. These two metrics complement the main results by further evaluating the success of target-label flipping and the degree of input text modification.

H.1 Hard Label Flipping Rate

Hard Label Flipping Rate (HLFR) measures validity, defined as the fraction of SCEs that successfully change the model prediction to the targeted label $\hat{y}$. The same model $\mathcal{M}$ is used both for generating SCE $\tilde{x}$ and for evaluating whether the predicted label is shifted. Formally, on N evaluation instances, HLFR is computed as:

$$\text{HLFR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\big[\mathcal{M}(\tilde{x}_i) = \hat{y}\big]$$

where $\mathbb{I}[\cdot]$ is an indicator function.

H.2 Textual Similarity

Textual Similarity (TS) measures the extent of modifications made to the original input x. Following Dehghanighobadi et al. (2025); Wang et al. (2025); Bhattacharjee et al. (2024), we employ a normalized Levenshtein distance d, which captures all applied edits. Formally, over N evaluation instances, TS is computed as:

$$\text{TS} = \frac{1}{N}\sum_{i=1}^{N}\frac{d(x_i, \tilde{x}_i)}{|x_i|}$$

where $d(x_i, \tilde{x}_i)$ is the Levenshtein distance between the original input x_i and its SCE $\tilde{x}_i$, and $|x_i|$ is the length of the original input used for normalization.

I The Impact of Direct Preference Optimization

I.1 Impact on Model General Capability

We explore the impact of DPO training by MACRO on general model capabilities. We focus on two perspectives – reasoning and multilingual capability – for which we employ the MMLU-ProX (Xuan et al., 2025) and MMLU (Hendrycks et al., 2021) datasets. MMLU is a widely used benchmark for measuring multitask language understanding across 57 subjects. In contrast, MMLU-ProX extends the more

Setting	Lang	SIB200					Taxi1500				
		SLFR ↑	HLFR ↑	PPL ↓	SS ↑	TS ↓	SLFR ↑	HLFR ↑	PPL ↓	SS ↑	TS ↓
w/ align	en	0.566 (↓0.171)	0.384 (↓0.172)	36.39 (↓1.48)	0.690 (↑0.027)	0.274 (↓0.034)	0.676 (↓0.045)	0.523 (↓0.009)	24.27 (↓0.78)	0.691 (↑0.001)	0.477 (↑0.009)
	de	0.697 (↓0.030)	0.596 (↑0.010)	26.48 (↓1.28)	0.687 (↓0.018)	0.280 (↑0.021)	0.748 (↑0.036)	0.477 (↓0.028)	23.07 (↑0.20)	0.688 (↑0.030)	0.498 (↓0.046)
	zh	0.697 (→0.000)	0.626 (→0.000)	34.73 (↓0.19)	0.668 (↓0.035)	0.589 (↑0.023)	0.667 (↑0.009)	0.405 (↑0.027)	34.20 (↑0.57)	0.707 (↑0.002)	0.846 (↓0.019)
	ar	0.636 (↓0.122)	0.525 (↓0.121)	29.72 (↓0.80)	0.626 (↓0.027)	0.295 (↑0.014)	0.775 (↑0.018)	0.441 (→0.000)	40.06 (↑2.26)	0.645 (↑0.019)	0.612 (↓0.026)
	vi	0.697 (↓0.020)	0.636 (↑0.080)	23.83 (↑0.72)	0.702 (↓0.005)	0.307 (↑0.003)	0.712 (↓0.072)	0.477 (↓0.037)	18.28 (↑0.30)	0.611 (↓0.026)	0.620 (↑0.019)
	sw	0.576 (↑0.010)	0.182 (↓0.081)	20.21 (↑0.48)	0.774 (↓0.019)	0.231 (↓0.524)	0.261 (↓0.009)	0.063 (↓0.018)	16.48 (↓0.30)	0.829 (→0.000)	0.776 (↓0.474)
	ru	0.687 (↑0.030)	0.535 (↑0.030)	21.29 (↓0.24)	0.684 (↓0.025)	0.306 (↑0.023)	0.559 (↓0.054)	0.378 (↓0.018)	20.10 (↓0.13)	0.718 (↓0.012)	0.593 (↑0.002)

Table 7: Ablation study with align reward.

Dataset		SIB200				TAXI1500			
Model	Language	HLFR (↑)		TS (↓)		HLFR (↑)		TS (↓)	
		DG-CF	MACRO	DG-CF	MACRO	DG-CF	MACRO	DG-CF	MACRO
Gemma3-4B	en	0.559	**0.576**	0.529	**0.427**	0.434	**0.514**	0.670	**0.606**
	de	0.499	**0.545**	0.389	**0.355**	0.405	**0.450**	0.580	**0.551**
	zh	0.629	**0.778**	**0.467**	0.666	**0.405**	0.369	**0.807**	0.961
	ar	0.548	**0.646**	0.401	**0.399**	0.342	**0.351**	0.716	**0.659**
	vi	0.508	**0.586**	0.456	**0.437**	**0.377**	0.369	0.631	**0.610**
	sw	0.549	**0.606**	**0.446**	0.473	0.357	**0.360**	0.647	**0.609**
	ru	0.514	**0.606**	0.459	**0.435**	0.394	**0.432**	**0.650**	0.666
Qwen3-4B	en	0.485	**0.618**	**0.251**	0.294	0.485	**0.505**	0.480	**0.439**
	de	0.544	**0.578**	0.276	**0.263**	0.284	**0.333**	0.444	**0.439**
	zh	**0.765**	0.662	**0.459**	0.574	0.309	**0.387**	**0.589**	0.769
	ar	0.569	**0.574**	**0.255**	0.310	0.342	**0.378**	0.615	**0.556**
	vi	0.627	**0.637**	**0.318**	0.337	**0.356**	0.351	0.485	**0.522**
	sw	0.127	**0.137**	0.434	**0.242**	**0.107**	0.090	**0.312**	0.320
	ru	0.373	**0.564**	**0.185**	0.273	0.315	**0.414**	0.526	**0.518**
Qwen3-8B	en	0.489	**0.556**	**0.231**	0.308	0.387	**0.532**	0.474	**0.468**
	de	0.568	**0.586**	0.263	**0.259**	0.342	**0.505**	0.562	**0.544**
	zh	0.534	**0.626**	**0.320**	0.566	0.360	**0.378**	**0.689**	0.865
	ar	0.505	**0.646**	**0.169**	0.281	0.387	**0.441**	**0.622**	0.638
	vi	**0.614**	0.556	0.308	**0.304**	0.405	**0.514**	0.621	**0.601**
	sw	0.244	**0.263**	**0.324**	0.755	0.009	**0.081**	**0.693**	1.250
	ru	**0.572**	0.505	0.295	**0.283**	0.270	**0.396**	**0.566**	0.591
Gemma3-12B	en	0.763	**0.788**	0.471	**0.377**	0.736	**0.757**	0.731	**0.657**
	de	0.703	**0.784**	0.472	**0.352**	0.678	**0.694**	0.686	**0.631**
	zh	0.789	**0.869**	**0.554**	0.709	0.605	**0.631**	**0.850**	0.973
	ar	0.789	**0.889**	0.467	**0.407**	**0.672**	0.631	0.699	**0.670**
	vi	0.709	**0.737**	0.508	**0.401**	0.650	**0.676**	0.707	**0.655**
	sw	0.733	**0.747**	0.446	**0.358**	0.578	**0.640**	0.653	**0.634**
	ru	0.691	**0.717**	0.494	**0.378**	0.643	**0.685**	0.743	**0.693**

Table 12: HLFR and TS automatic evaluation results on SIB200 and TAXI1500. **Boldface** denotes the stronger method per language and metric between **DG-CF** and **MACRO**, while the wavy underline highlights the best score among the seven languages for the same model, method, and metric.

challenging MMLU-Pro benchmark, which features reasoning-focused questions and ten answer choices instead of four.

Figures 11, 12 and 13 demonstrate that, in general, models after applying MACRO can largely maintain or even enhance their reasoning and multilingual capabilities. Table 13 further confirms the same trend with overall performance: MACRO (DPO) consistently preserves or slightly improves performance relative to the base model on both

	MMLU			MMLU-ProX		
Model	Base	SFT	DPO	Base	SFT	DPO
Qwen3-4B	79.38	79.33	**79.78**	54.03	55.10	**55.37**
Gemma3-4B	67.13	64.88	**67.26**	43.38	43.45	**44.47**

Table 13: Overall accuracy (%) on MMLU and selected MMLU-ProX categories for the base, SFT, and DPO.

MMLU and MMLU-ProX. In contrast, SFT does not show the same level of consistency, especially for Gemma3-4B on MMLU. This is crucial, as we aim to improve the quality of self-generated counterfactual explanations while keeping general capabilities intact.

I.2 Impact on Cross-lingual Generalizability

To further assess whether the benefits of preference alignment transfer beyond the training languages, we evaluate Gemma3-4B and Qwen3-4B on six unseen languages: Bulgarian, Greek, Spanish, Hindi, Thai, and Turkish. This analysis complements our in-domain findings, where MACRO generally improves counterfactual quality and shows cross-lingual generalizability.

In general, Table 17 suggests that MACRO generalizes beyond the languages included in the training set, but its gains are more consistent on *validity* than on *fluency* or *minimality*. In other words, MACRO appears to learn a more transferable notion of *what kinds of edits are effective* for flipping predictions, whereas *how naturally and minimally those edits are performed* remains more sensitive to model family and dataset.

J Trade-off

Figure 14 presents the *validity-minimality* trade-off across all models and languages on SIB200 and Taxi1500.

K Translation-based Counterfactual

Table 14 illustrates the automatic evaluation results of translation-based counterfactuals generated by TB-CF. Although TB-CF yields higher *validity* than those of DG-CF or MACRO, they require excessive modifications. These drastic edits, driven by translation artifacts and cross-linguistic syntactic differences, cause TB-CF to drift too far from the source text. Consequently, they fail to isolate the specific features that influence model decisions and should no longer be viewed as reliable counterfactual explanations.

L Cross-lingual Edit Similarity

Figure 15a shows the cross-lingual edit similarity scores across all language pairs on Taxi1500. The trend of Taxi1500 is consistent with that of SIB200 (Figure 15b, §5.3). The cross-lingual edit consistency between Swahili and other languages is noticeably lower than that of other language pairs. Generally, after applying MACRO, the cross-lingual edit similarity increases.

M Ablation Study

M.1 Ablation Examples

We present qualitative multilingual counterfactual examples for SIB200 and TAXI1500 in Figures 10 and 11, respectively.

M.2 Ablation Results

We report ablation study results to analyze the role of each reward component in MACRO. Specifically, we remove $\mathcal{R}_{\text{flip}}$, $\mathcal{R}_{\text{aug}}$, and $\mathcal{R}_{\text{edit}}$ one at a time and compare the resulting models against the full reward setting. Table 15 and Table 16 report the automatic evaluation results on SIB200 and Taxi1500 using Qwen3-8B and Gemma3-4B. Overall, removing $\mathcal{R}_{\text{flip}}$ and $\mathcal{R}_{\text{aug}}$ most consistently reduces flip rates, while removing $\mathcal{R}_{\text{edit}}$ increases flip success at the cost of larger edits and lower semantic similarity.

M.3 Additional Analysis

To systematically compare how each language responds to component removal, based on Table 15, we compute the *relative degradation* for each metric as $\Delta_{\text{rel}} = \delta / v_{\text{full}}$, where δ is the absolute change reported in parentheses and v_{full} is the corresponding full-model value. We then aggregate these relative changes into three dimensions, i.e., *validity* (average across SLFR and HLFR (Appendix H.1)), *fluency* (PPL), and *minimality* (average across SS and TS (Appendix H.2)), averaged over both datasets. A positive value indicates degradation (validity dropped, PPL increased, or minimality worsened), while a negative value indicates improvement. Tables 18-19 present the per-setting results, and Table 21 shows the overall average across all three settings.

M.3.1 Language-Specific Observation

Analysis of w/o $\mathcal{R}_{\text{flip}}$ (Table 18). Removing the flip reward degrades *validity* across all languages,

Dataset		SIB200											
Model	Lang-uage	SLFR (↑)		HLFR (↑)		PPL (↓)		SS (↑)		TS (↓)		$\mathcal{R}_{\text{edit}}$ (↑)	
		DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF
Gemma3-4B	en	0.627	-	0.559	-	48.76	-	0.613	-	0.529	-	0.471	-
	de	0.592	**0.737** (↑ 0.145)	**0.499**	0.414 (↓ 0.085)	28.54	**11.95** (↓ 16.59)	**0.754**	0.267 (↓ 0.487)	**0.389**	7.640 (↑ 7.251)	**0.611**	-6.640 (↓ 7.251)
	zh	0.662	**0.687** (↑ 0.025)	**0.629**	0.313 (↓ 0.316)	35.09	**15.68** (↓ 19.41)	**0.692**	0.074 (↓ 0.618)	**0.467**	22.770 (↑ 22.303)	**0.533**	-21.770 (↓ 22.303)
	ar	0.609	**0.657** (↑ 0.048)	**0.548**	0.293 (↓ 0.255)	27.74	**19.29** (↓ 8.45)	**0.663**	0.073 (↓ 0.590)	**0.401**	8.045 (↑ 7.644)	**0.599**	-7.045 (↓ 7.644)
	vi	0.590	**0.677** (↑ 0.087)	**0.508**	0.434 (↓ 0.074)	18.88	**10.80** (↓ 8.08)	**0.711**	0.317 (↓ 0.394)	**0.456**	7.959 (↑ 7.503)	**0.544**	-6.959 (↓ 7.503)
	sw	0.676	**0.687** (↑ 0.011)	**0.549**	0.424 (↓ 0.125)	18.71	**13.59** (↓ 5.12)	**0.681**	0.427 (↓ 0.254)	**0.446**	7.620 (↑ 7.174)	**0.554**	-6.620 (↓ 7.174)
	ru	0.605	**0.677** (↑ 0.072)	**0.514**	0.404 (↓ 0.110)	20.26	**12.04** (↓ 8.22)	**0.737**	0.288 (↓ 0.449)	**0.459**	8.039 (↑ 7.580)	**0.541**	-7.039 (↓ 7.580)
Qwen3-4B	en	0.531	-	0.482	-	58.52	-	0.801	-	0.224	-	0.776	-
	de	**0.709**	0.596 (↓ 0.113)	**0.612**	0.455 (↓ 0.157)	45.15	**41.40** (↓ 3.75)	0.658	**0.733** (↑ 0.075)	**0.306**	0.562 (↑ 0.256)	**0.694**	0.438 (↓ 0.256)
	zh	0.767	**0.768** (↑ 0.001)	**0.724**	0.374 (↓ 0.350)	**43.75**	44.39 (↑ 0.64)	**0.598**	0.078 (↓ 0.520)	**0.438**	1.211 (↑ 0.773)	**0.562**	-0.211 (↓ 0.773)
	ar	0.646	**0.768** (↑ 0.122)	**0.591**	0.374 (↓ 0.217)	40.77	**40.49** (↓ 0.28)	**0.704**	0.084 (↓ 0.620)	**0.251**	0.941 (↑ 0.690)	**0.749**	0.059 (↓ 0.690)
	vi	**0.678**	0.475 (↓ 0.203)	**0.578**	0.394 (↓ 0.184)	26.15	**20.38** (↓ 5.77)	0.649	**0.733** (↑ 0.084)	**0.320**	0.579 (↑ 0.259)	**0.680**	0.421 (↓ 0.259)
	sw	0.512	**0.566** (↑ 0.054)	0.137	**0.141** (↑ 0.004)	42.85	**18.69** (↓ 24.16)	**0.714**	0.491 (↓ 0.223)	**0.397**	1.889 (↑ 1.492)	**0.603**	-0.889 (↓ 1.492)
	ru	0.496	**0.505** (↑ 0.009)	0.423	**0.444** (↑ 0.021)	**23.73**	32.23 (↑ 8.50)	**0.793**	0.731 (↓ 0.062)	**0.199**	0.612 (↑ 0.413)	**0.801**	0.388 (↓ 0.413)
Qwen3-8B	en	0.594	-	0.489	-	53.24	-	0.731	-	0.231	-	0.769	-
	de	**0.652**	0.596 (↓ 0.056)	**0.568**	0.495 (↓ 0.073)	32.09	**30.57** (↓ 1.52)	**0.686**	0.661 (↓ 0.025)	**0.263**	0.558 (↑ 0.295)	**0.737**	0.442 (↓ 0.295)
	zh	0.615	**0.717** (↑ 0.102)	**0.534**	0.404 (↓ 0.130)	38.59	**34.98** (↓ 3.61)	**0.681**	0.083 (↓ 0.598)	**0.320**	1.235 (↑ 0.915)	**0.680**	-0.235 (↓ 0.915)
	ar	0.576	**0.758** (↑ 0.182)	**0.505**	0.404 (↓ 0.101)	**34.88**	34.94 (↑ 0.06)	**0.735**	0.080 (↓ 0.655)	**0.169**	0.951 (↑ 0.782)	**0.831**	0.049 (↓ 0.782)
	vi	**0.681**	0.586 (↓ 0.095)	**0.614**	0.515 (↓ 0.099)	22.66	**19.40** (↓ 3.26)	**0.692**	0.656 (↓ 0.036)	**0.308**	0.583 (↑ 0.275)	**0.692**	0.417 (↓ 0.275)
	sw	0.533	**0.556** (↑ 0.023)	**0.244**	0.242 (↓ 0.002)	22.93	**22.40** (↓ 0.53)	**0.844**	0.521 (↓ 0.323)	**0.324**	1.512 (↑ 1.188)	**0.676**	-0.512 (↓ 1.188)
	ru	**0.635**	0.596 (↓ 0.039)	**0.572**	0.455 (↓ 0.117)	**22.08**	22.43 (↑ 0.35)	**0.696**	0.662 (↓ 0.034)	**0.295**	0.609 (↑ 0.314)	**0.705**	0.391 (↓ 0.314)
Gemma3-12B	en	0.830	-	0.763	-	53.90	-	0.530	-	0.471	-	0.529	-
	de	0.767	**0.788** (↑ 0.021)	**0.703**	0.667 (↓ 0.036)	28.13	**13.32** (↓ 14.81)	**0.574**	0.239 (↓ 0.335)	**0.472**	7.754 (↑ 7.282)	**0.528**	-6.754 (↓ 7.282)
	zh	**0.842**	0.828 (↓ 0.014)	**0.789**	0.576 (↓ 0.213)	31.61	**16.96** (↓ 14.65)	**0.551**	0.073 (↓ 0.478)	**0.554**	22.704 (↑ 22.150)	**0.446**	-21.704 (↓ 22.150)
	ar	**0.867**	0.818 (↓ 0.049)	**0.789**	0.556 (↓ 0.233)	25.48	**25.19** (↓ 0.29)	**0.521**	0.063 (↓ 0.458)	**0.467**	8.081 (↑ 7.614)	**0.533**	-7.081 (↓ 7.614)
	vi	**0.794**	0.768 (↓ 0.026)	**0.709**	0.657 (↓ 0.052)	18.54	**12.51** (↓ 6.03)	**0.557**	0.255 (↓ 0.302)	**0.508**	8.144 (↑ 7.636)	**0.492**	-7.144 (↓ 7.636)
	sw	0.814	**0.828** (↑ 0.014)	**0.733**	0.697 (↓ 0.036)	14.86	**12.45** (↓ 2.41)	**0.653**	0.399 (↓ 0.254)	**0.446**	7.464 (↑ 7.018)	**0.554**	-6.464 (↓ 7.018)
	ru	**0.779**	0.758 (↓ 0.021)	**0.691**	0.626 (↓ 0.065)	21.03	**15.23** (↓ 5.80)	**0.585**	0.240 (↓ 0.345)	**0.494**	7.769 (↑ 7.275)	**0.506**	-6.769 (↓ 7.275)
Dataset		TAXI1500											
Model	Lang-uage	SLFR (↑)		HLFR (↑)		PPL (↓)		SS (↑)		TS (↓)		$\mathcal{R}_{\text{edit}}$ (↑)	
		DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF	DG-CF	TB-CF
Gemma3-4B	en	0.531	-	0.434	-	36.24	-	0.507	-	0.670	-	0.330	-
	de	0.557	**0.874** (↑ 0.317)	**0.405**	0.324 (↓ 0.081)	33.93	**11.59** (↓ 22.34)	**0.639**	0.293 (↓ 0.346)	**0.580**	8.018 (↑ 7.438)	**0.420**	-7.018 (↓ 7.438)
	zh	0.523	**0.892** (↑ 0.369)	**0.405**	0.315 (↓ 0.090)	39.48	**14.64** (↓ 24.84)	**0.612**	0.291 (↓ 0.421)	**0.807**	26.717 (↑ 25.910)	**0.193**	-25.717 (↓ 25.910)
	ar	0.522	**0.919** (↑ 0.397)	**0.342**	0.333 (↓ 0.009)	41.88	**17.79** (↓ 24.09)	**0.628**	0.327 (↓ 0.301)	**0.716**	11.583 (↑ 10.867)	**0.284**	-10.583 (↓ 10.867)
	vi	0.504	**0.838** (↑ 0.334)	**0.377**	0.270 (↓ 0.107)	18.47	**10.46** (↓ 8.01)	**0.613**	0.295 (↓ 0.318)	**0.631**	7.942 (↑ 7.311)	**0.369**	-6.942 (↓ 7.311)
	sw	0.569	**0.865** (↑ 0.296)	**0.357**	0.351 (↓ 0.006)	25.59	**14.78** (↓ 10.81)	**0.670**	0.503 (↓ 0.167)	**0.647**	7.537 (↑ 6.890)	**0.353**	-6.537 (↓ 6.890)
	ru	0.532	**0.865** (↑ 0.333)	**0.394**	0.288 (↓ 0.106)	21.49	**12.73** (↓ 8.76)	**0.656**	0.317 (↓ 0.339)	**0.650**	9.932 (↑ 9.282)	**0.350**	-8.932 (↓ 9.282)
Qwen3-4B	en	0.623	-	0.485	-	32.64	-	0.614	-	0.480	-	0.520	-
	de	**0.486**	**0.486** (→ 0.000)	0.284	**0.342** (↑ 0.058)	40.17	**37.80** (↓ 2.37)	**0.662**	0.589 (↓ 0.073)	**0.444**	0.731 (↑ 0.287)	**0.556**	0.269 (↓ 0.287)
	zh	0.587	**0.622** (↑ 0.035)	**0.309**	0.306 (↓ 0.003)	45.84	**35.72** (↓ 10.12)	**0.674**	0.600 (↓ 0.074)	**0.589**	0.927 (↑ 0.338)	**0.411**	0.073 (↓ 0.338)
	ar	**0.681**	0.649 (↓ 0.032)	**0.342**	0.279 (↓ 0.063)	86.15	**55.98** (↓ 30.17)	**0.577**	0.563 (↓ 0.014)	**0.615**	0.839 (↑ 0.224)	**0.385**	0.161 (↓ 0.224)
	vi	0.651	**0.676** (↑ 0.025)	**0.356**	0.288 (↓ 0.068)	25.67	**19.18** (↓ 6.49)	**0.660**	0.565 (↓ 0.095)	**0.485**	0.814 (↑ 0.329)	**0.515**	0.186 (↓ 0.329)
	sw	0.463	**0.505** (↑ 0.042)	0.107	**0.117** (↑ 0.010)	25.61	**14.90** (↓ 10.71)	**0.840**	0.553 (↓ 0.287)	**0.312**	1.765 (↑ 1.453)	**0.688**	-0.765 (↓ 1.453)
	ru	0.529	**0.631** (↑ 0.102)	0.315	**0.324** (↑ 0.009)	27.73	**25.04** (↓ 2.69)	**0.675**	0.620 (↓ 0.055)	**0.526**	0.848 (↑ 0.322)	**0.474**	0.152 (↓ 0.322)
Qwen3-8B	en	0.559	-	0.387	-	30.95	-	0.682	-	0.474	-	0.526	-
	de	**0.495**	0.432 (↓ 0.063)	**0.342**	0.252 (↓ 0.090)	**25.85**	26.46 (↑ 0.61)	0.617	**0.646** (↑ 0.029)	**0.562**	0.727 (↑ 0.165)	**0.438**	0.273 (↓ 0.165)
	zh	**0.631**	0.577 (↓ 0.054)	**0.360**	0.324 (↓ 0.036)	38.92	**33.37** (↓ 5.55)	**0.662**	0.639 (↓ 0.023)	**0.689**	0.933 (↑ 0.244)	**0.311**	0.067 (↓ 0.244)
	ar	0.658	**0.667** (↑ 0.009)	**0.387**	0.333 (↓ 0.054)	50.29	**41.30** (↓ 8.99)	**0.639**	0.600 (↓ 0.039)	**0.622**	0.808 (↑ 0.186)	**0.378**	0.192 (↓ 0.186)
	vi	**0.577**	0.514 (↓ 0.063)	**0.405**	0.306 (↓ 0.099)	18.83	**16.36** (↓ 2.47)	**0.613**	0.603 (↓ 0.010)	**0.621**	0.849 (↑ 0.228)	**0.379**	0.151 (↓ 0.228)
	sw	0.108	**0.180** (↑ 0.072)	0.009	**0.054** (↑ 0.045)	18.96	**18.90** (↓ 0.06)	**0.833**	0.610 (↓ 0.223)	**0.693**	1.241 (↑ 0.548)	**0.307**	-0.241 (↓ 0.548)
	ru	**0.450**	0.378 (↓ 0.072)	0.270	**0.288** (↑ 0.018)	18.73	**18.33** (↓ 0.40)	**0.676**	0.671 (↓ 0.005)	**0.566**	0.847 (↑ 0.281)	**0.434**	0.153 (↓ 0.281)
Gemma3-12B	en	0.832	-	0.736	-	41.48	-	0.437	-	0.731	-	0.269	-
	de	0.791	**0.937** (↑ 0.146)	**0.678**	0.667 (↓ 0.011)	26.50	**13.01** (↓ 13.49)	**0.537**	0.257 (↓ 0.280)	**0.686**	8.387 (↑ 7.701)	**0.314**	-7.387 (↓ 7.701)
	zh	0.682	**0.838** (↑ 0.156)	0.605	**0.640** (↑ 0.035)	27.90	**17.40** (↓ 10.50)	**0.544**	0.257 (↓ 0.287)	**0.850**	26.798 (↑ 25.948)	**0.150**	-25.798 (↓ 25.948)
	ar	0.777	**0.802** (↑ 0.025)	**0.672**	0.631 (↓ 0.041)	37.25	**22.22** (↓ 15.03)	**0.537**	0.298 (↓ 0.239)	**0.699**	11.946 (↑ 11.247)	**0.301**	-10.946 (↓ 11.247)
	vi	0.744	**0.892** (↑ 0.148)	**0.650**	0.559 (↓ 0.091)	16.18	**12.08** (↓ 4.10)	**0.517**	0.247 (↓ 0.270)	**0.707**	8.459 (↑ 7.752)	**0.293**	-7.459 (↓ 7.752)
	sw	0.694	**0.919** (↑ 0.225)	0.578	**0.766** (↑ 0.188)	17.46	**13.54** (↓ 3.92)	**0.684**	0.476 (↓ 0.208)	**0.653**	7.483 (↑ 6.830)	**0.347**	-6.483 (↓ 6.830)
	ru	0.749	**0.883** (↑ 0.134)	**0.643**	0.604 (↓ 0.039)	20.49	**15.75** (↓ 4.74)	**0.600**	0.271 (↓ 0.329)	**0.743**	9.984 (↑ 9.241)	**0.257**	-8.984 (↓ 9.241)

Table 14: Automatic evaluation on SIB200 and TAXI1500. We report **DG-CF** and **TB-CF** on SLFR, HLFR, PPL, SS, TS, and $\mathcal{R}_{\text{edit}}$. **Boldface** denotes the better method for each language and metric.

with English suffering the most (+19.18%), followed by Chinese and German. Vietnamese and Swahili are the least affected (+6.43% and +7.88%). *Minimality* improves modestly for most languages, but Swahili shows a dramatic degradation of 36.27%, which reveals that the full model's Swahili counterfactuals involve relatively large edit distances, and without flip pressure, the model collapses to near-paraphrases.

Analysis of w/o $\mathcal{R}_{\text{aug}}$ (Table 19). The augmentation reward primarily affects *validity*, with all languages showing consistent degradation. English, German, and Swahili are the most affected. Russian is the least affected (+4.81%), suggesting that its counterfactuals already cross the decision boundary with sufficient margin without augmentation. Minimality is moderately degraded. Arabic again stands out with +8.34% fluency degradation, further evidencing its unique sensitivity to any modification of the reward structure. Swahili again shows the anomalous minimality deterioration (−37.42%).

Setting	Lang	SIB200					Taxi1500				
		SLFR ↑	HLFR ↑	PPL ↓	SS ↑	TS ↓	SLFR ↑	HLFR ↑	PPL ↓	SS ↑	TS ↓
w/o $\mathcal{R}_{\text{flip}}$	en	0.525 (↓0.212)	0.384 (↓0.172)	35.95 (↓1.92)	0.705 (↑0.042)	0.261 (↓0.047)	0.685 (↓0.036)	0.468 (↓0.064)	24.65 (↓0.40)	0.707 (↑0.017)	0.452 (↓0.016)
	de	0.586 (↓0.141)	0.485 (↓0.101)	27.86 (↑0.10)	0.698 (↓0.007)	0.260 (↑0.001)	0.712 (→0.000)	0.414 (↓0.091)	24.70 (↑1.83)	0.706 (↑0.048)	0.493 (↓0.051)
	zh	0.576 (↓0.121)	0.485 (↓0.141)	35.48 (↑0.56)	0.694 (↓0.009)	0.558 (↓0.008)	0.640 (↓0.018)	0.405 (↑0.027)	35.83 (↑2.20)	0.752 (↑0.047)	0.819 (↓0.046)
	ar	0.657 (↓0.101)	0.495 (↓0.151)	30.31 (↓0.21)	0.668 (↑0.015)	0.270 (↓0.011)	0.757 (→0.000)	0.459 (↑0.018)	47.18 (↑9.38)	0.658 (↑0.032)	0.566 (↓0.072)
	vi	0.657 (↓0.060)	0.485 (↓0.071)	23.59 (↑0.48)	0.707 (→0.000)	0.284 (↓0.020)	0.748 (↓0.036)	0.495 (↓0.019)	19.00 (↑1.02)	0.675 (↑0.038)	0.562 (↓0.039)
	sw	0.515 (↓0.051)	0.212 (↓0.051)	20.22 (↑0.49)	0.794 (↑0.001)	0.217 (↓0.538)	0.279 (↑0.009)	0.081 (→0.000)	17.10 (↑0.32)	0.841 (↑0.012)	0.347 (↓0.903)
	ru	0.697 (↑0.040)	0.566 (↑0.061)	21.77 (↑0.24)	0.705 (↓0.004)	0.286 (↑0.003)	0.550 (↓0.063)	0.351 (↓0.045)	20.91 (↑0.68)	0.758 (↑0.028)	0.541 (↓0.050)
w/o $\mathcal{R}_{\text{aug}}$	en	0.626 (↓0.111)	0.455 (↓0.101)	35.03 (↓2.84)	0.702 (↑0.039)	0.263 (↓0.045)	0.676 (↓0.045)	0.477 (↓0.055)	25.13 (↑0.08)	0.730 (↑0.040)	0.435 (↓0.033)
	de	0.636 (↓0.091)	0.545 (↓0.041)	27.17 (↓0.59)	0.691 (↓0.014)	0.266 (↑0.007)	0.739 (↑0.027)	0.360 (↓0.145)	23.92 (↑1.05)	0.711 (↑0.053)	0.475 (↓0.069)
	zh	0.646 (↓0.051)	0.576 (↓0.050)	34.73 (↓0.19)	0.670 (↓0.033)	0.586 (↑0.020)	0.595 (↓0.063)	0.315 (↓0.063)	36.91 (↑3.28)	0.747 (↑0.042)	0.822 (↓0.043)
	ar	0.677 (↓0.081)	0.566 (↓0.080)	29.07 (↓1.45)	0.672 (↑0.019)	0.285 (↑0.004)	0.703 (↓0.054)	0.423 (↓0.018)	42.12 (↑4.32)	0.674 (↑0.048)	0.572 (↓0.066)
	vi	0.657 (↓0.060)	0.576 (↑0.020)	23.33 (↑0.22)	0.682 (↓0.025)	0.302 (↓0.002)	0.721 (↓0.063)	0.432 (↓0.082)	18.90 (↑0.92)	0.653 (↑0.016)	0.576 (↓0.025)
	sw	0.525 (↓0.041)	0.202 (↓0.061)	19.30 (↓0.43)	0.783 (↓0.010)	0.214 (↓0.541)	0.297 (↑0.027)	0.072 (↓0.009)	17.01 (↑0.23)	0.859 (↑0.030)	0.336 (↓0.914)
	ru	0.646 (↓0.011)	0.525 (↑0.020)	21.07 (↓0.46)	0.703 (↓0.006)	0.287 (↑0.004)	0.631 (↑0.018)	0.369 (↓0.027)	22.08 (↑1.85)	0.743 (↑0.013)	0.532 (↓0.059)
w/o $\mathcal{R}_{\text{edit}}$	en	0.929 (↑0.192)	0.818 (↑0.262)	28.91 (↓8.96)	0.348 (↓0.315)	0.665 (↑0.357)	0.847 (↑0.126)	0.667 (↑0.135)	17.99 (↓7.06)	0.431 (↓0.259)	0.748 (↑0.280)
	de	0.929 (↑0.202)	0.828 (↑0.242)	21.06 (↓6.70)	0.395 (↓0.310)	0.626 (↑0.367)	0.856 (↑0.144)	0.712 (↑0.207)	17.14 (↓5.73)	0.431 (↓0.227)	0.781 (↑0.237)
	zh	0.939 (↑0.242)	0.909 (↑0.283)	26.51 (↓8.41)	0.398 (↓0.305)	0.956 (↑0.390)	0.793 (↑0.135)	0.613 (↑0.235)	20.54 (↓13.09)	0.442 (↓0.263)	1.218 (↑0.353)
	ar	0.929 (↑0.171)	0.848 (↑0.202)	21.87 (↓8.65)	0.331 (↓0.322)	0.621 (↑0.340)	0.910 (↑0.153)	0.658 (↑0.217)	25.11 (↓12.69)	0.416 (↓0.210)	0.945 (↑0.307)
	vi	0.899 (↑0.182)	0.869 (↑0.313)	15.49 (↓7.62)	0.393 (↓0.314)	0.667 (↑0.363)	0.847 (↑0.063)	0.640 (↑0.126)	12.22 (↓5.76)	0.399 (↓0.238)	0.826 (↑0.225)
	sw	0.717 (↑0.151)	0.303 (↑0.040)	21.60 (↑1.87)	0.633 (↓0.160)	0.510 (↓0.245)	0.387 (↑0.117)	0.189 (↑0.108)	18.41 (↑1.63)	0.727 (↓0.102)	0.755 (↓0.495)
	ru	0.960 (↑0.303)	0.848 (↑0.343)	16.50 (↓5.03)	0.370 (↓0.339)	0.661 (↑0.378)	0.757 (↑0.144)	0.568 (↑0.172)	13.83 (↓6.40)	0.457 (↓0.273)	0.927 (↑0.336)

Table 15: Automatic evaluation results for different reward ablations on SIB200 and Taxi1500 using `Qwen3-8B`. Each cell reports the absolute metric value, and the number in parentheses indicates the absolute change relative to the full reward MACRO configuration.

Setting	Lang	SIB200					Taxi1500				
		SLFR (↑)	HLFR (↑)	PPL (↓)	SS (↑)	TS (↓)	SLFR (↑)	HLFR (↑)	PPL (↓)	SS (↑)	TS (↓)
w/o $\mathcal{R}_{\text{flip}}$	en	0.588 (↓0.048)	0.505 (↓0.071)	32.81 (↑1.08)	0.679 (↑0.029)	0.381 (↓0.046)	0.532 (↓0.063)	0.441 (↓0.073)	23.69 (↑0.15)	0.608 (↑0.021)	0.593 (↓0.013)
	de	0.515 (↓0.121)	0.436 (↓0.109)	25.29 (↑0.85)	0.783 (↑0.008)	0.317 (↓0.038)	0.586 (↓0.072)	0.423 (↓0.027)	24.86 (↑1.82)	0.713 (↓0.012)	0.539 (↓0.012)
	zh	0.632 (↓0.166)	0.559 (↓0.219)	31.30 (↑1.23)	0.740 (↑0.084)	0.608 (↓0.058)	0.532 (↓0.072)	0.351 (↓0.018)	29.89 (↓2.53)	0.700 (↑0.030)	0.943 (↓0.018)
	ar	0.647 (↓0.010)	0.559 (↓0.087)	28.18 (↑1.93)	0.698 (↑0.048)	0.362 (↓0.037)	0.486 (↓0.082)	0.306 (↓0.045)	35.27 (↑4.78)	0.667 (↑0.006)	0.627 (↓0.032)
	vi	0.539 (↓0.107)	0.456 (↓0.130)	19.34 (↑0.64)	0.754 (↑0.040)	0.391 (↓0.046)	0.414 (↓0.091)	0.333 (↓0.036)	17.48 (↑1.17)	0.679 (↑0.018)	0.576 (↓0.034)
	sw	0.623 (↓0.114)	0.520 (↓0.086)	17.55 (↓0.04)	0.746 (↑0.078)	0.379 (↓0.094)	0.622 (↓0.018)	0.405 (↑0.045)	18.15 (↓1.72)	0.732 (↓0.024)	0.637 (↑0.028)
	ru	0.574 (↓0.062)	0.475 (↓0.131)	20.10 (↑0.96)	0.778 (↑0.040)	0.348 (↓0.087)	0.514 (↓0.090)	0.369 (↓0.063)	19.15 (↑0.20)	0.735 (↑0.053)	0.622 (↓0.044)
w/o $\mathcal{R}_{\text{aug}}$	en	0.564 (↓0.072)	0.490 (↓0.086)	32.41 (↑0.68)	0.663 (↑0.013)	0.395 (↓0.032)	0.577 (↓0.018)	0.486 (↓0.028)	23.56 (↑0.02)	0.613 (↑0.026)	0.584 (↓0.022)
	de	0.564 (↓0.072)	0.456 (↓0.089)	25.55 (↑1.11)	0.753 (↓0.022)	0.366 (↑0.011)	0.622 (↓0.036)	0.459 (↑0.009)	23.62 (↑0.58)	0.721 (↓0.004)	0.530 (↓0.021)
	zh	0.701 (↓0.097)	0.642 (↓0.136)	30.12 (↑0.05)	0.721 (↑0.065)	0.672 (↑0.006)	0.595 (↓0.009)	0.432 (↑0.063)	32.68 (↑0.26)	0.693 (↑0.023)	0.940 (↓0.021)
	ar	0.642 (↓0.015)	0.549 (↓0.097)	27.38 (↑1.13)	0.694 (↑0.044)	0.382 (↓0.017)	0.541 (↓0.027)	0.369 (↑0.018)	33.00 (↑2.51)	0.662 (↑0.001)	0.659 (→0.000)
	vi	0.574 (↓0.072)	0.510 (↓0.076)	18.85 (↑0.15)	0.741 (↑0.027)	0.420 (↓0.017)	0.486 (↓0.019)	0.396 (↑0.027)	17.32 (↑1.01)	0.662 (↑0.001)	0.593 (↓0.017)
	sw	0.730 (↓0.007)	0.623 (↑0.017)	17.89 (↑0.30)	0.728 (↑0.060)	0.397 (↓0.076)	0.541 (↓0.099)	0.324 (↓0.036)	17.43 (↓2.44)	0.749 (↓0.007)	0.622 (↑0.013)
	ru	0.588 (↓0.048)	0.515 (↓0.091)	20.30 (↑1.16)	0.771 (↑0.033)	0.382 (↓0.053)	0.523 (↓0.081)	0.414 (↓0.018)	19.27 (↑0.32)	0.712 (↑0.030)	0.632 (↓0.034)
w/o $\mathcal{R}_{\text{edit}}$	en	0.789 (↑0.153)	0.750 (↑0.174)	29.29 (↓2.44)	0.506 (↓0.144)	0.689 (↑0.262)	0.676 (↑0.081)	0.559 (↑0.045)	22.51 (↓1.03)	0.441 (↓0.146)	0.732 (↑0.126)
	de	0.730 (↑0.094)	0.686 (↑0.141)	22.97 (↓1.47)	0.667 (↓0.108)	0.538 (↑0.183)	0.676 (↑0.018)	0.505 (↑0.055)	23.45 (↑0.41)	0.585 (↓0.140)	0.686 (↑0.135)
	zh	0.755 (↓0.043)	0.730 (↓0.048)	27.40 (↓2.67)	0.634 (↓0.022)	0.830 (↑0.164)	0.676 (↑0.072)	0.514 (↑0.145)	23.94 (↓8.48)	0.552 (↓0.118)	1.367 (↑0.406)
	ar	0.799 (↑0.142)	0.725 (↑0.079)	24.34 (↓1.91)	0.567 (↓0.083)	0.576 (↑0.177)	0.703 (↑0.135)	0.523 (↑0.172)	27.42 (↓3.07)	0.568 (↓0.093)	0.926 (↑0.267)
	vi	0.716 (↑0.070)	0.672 (↑0.086)	15.77 (↓2.93)	0.633 (↓0.081)	0.626 (↑0.189)	0.568 (↑0.063)	0.468 (↑0.099)	14.56 (↓1.75)	0.557 (↓0.104)	0.733 (↑0.123)
	sw	0.833 (↑0.096)	0.721 (↑0.115)	16.17 (↓1.42)	0.672 (↑0.004)	0.554 (↑0.081)	0.658 (↑0.018)	0.432 (↑0.072)	16.99 (↓2.88)	0.684 (↓0.072)	0.816 (↑0.207)
	ru	0.770 (↑0.134)	0.706 (↑0.100)	17.75 (↓1.39)	0.660 (↓0.078)	0.607 (↑0.172)	0.631 (↑0.027)	0.514 (↑0.082)	16.73 (↓2.22)	0.624 (↓0.058)	0.781 (↑0.115)

Table 16: Automatic evaluation results for different reward ablations on SIB200 and Taxi1500 using `Gemma3-4B`. Each cell reports the absolute metric value, and the number in parentheses indicates the absolute change relative to the full reward MACRO configuration.

Dataset		SIB200						TAXI1500					
Model	Language	SLFR (↑)		PPL (↓)		SS (↑)		SLFR (↑)		PPL (↓)		SS (↑)	
		DG-CF	MACRO	DG-CF	MACRO	DG-CF	MACRO	DG-CF	MACRO	DG-CF	MACRO	DG-CF	MACRO
Gemma3-4B	bg	0.632	0.627	14.78	15.61	0.722	0.768	-	-	-	-	-	-
	el	0.627	0.672	9.65	10.69	0.691	0.685	-	-	-	-	-	-
	es	0.588	0.559	21.17	23.92	0.709	0.750	0.595	0.622	19.33	19.00	0.658	0.702
	hi	0.676	0.740	4.52	4.77	0.682	0.681	0.640	0.649	5.92	5.99	0.635	0.719
	th	0.618	0.637	6.60	7.26	0.674	0.683	0.595	0.631	8.94	8.57	0.574	0.636
	tr	0.642	0.657	16.20	18.68	0.676	0.710	-	-	-	-	-	-
Qwen3-4B	bg	0.446	0.657	17.52	19.36	0.859	0.813	-	-	-	-	-	-
	el	0.446	0.515	12.41	12.55	0.806	0.800	-	-	-	-	-	-
	es	0.544	0.711	23.53	26.21	0.793	0.740	0.505	0.586	19.05	21.93	0.665	0.743
	hi	0.412	0.534	5.22	5.32	0.822	0.792	0.387	0.270	6.41	6.94	0.676	0.793
	th	0.583	0.613	7.98	8.31	0.748	0.703	0.595	0.622	8.97	10.00	0.619	0.688
	tr	0.583	0.632	20.70	22.75	0.762	0.736	-	-	-	-	-	-

Table 17: Zero-shot evaluation results on SIB200 and TAXI1500 for languages unseen during training. We compare DG-CF and MACRO across three metrics: soft label flipping rate (SLFR (↑)), perplexity (PPL (↓)), and semantic similarity (SS (↑)). **best** denote the best results for each language-metric pair, respectively.

Analysis of w/o $\mathcal{R}_{\text{edit}}$ (Table 20). This setting produces the most extreme degradation patterns. *Validity* improves substantially for all languages (the edit constraint was *restraining* flip success), with Swahili showing the largest gain (54.64%)—its SLFR and HLFR roughly double without the

Lang	Validity	Fluency	Minimality
en	+19.18	−3.33	−6.87
de	+14.28	−3.63	−4.52
zh	+14.43	−3.59	−3.68
ar	+10.12	+12.06	−5.65
vi	+6.43	−3.55	−4.76
sw	+7.88	−2.10	−36.27
ru	+9.20	−2.12	−3.48

Table 18: Relative degradation (%) under w/o $\mathcal{R}_{\text{flip}}$.

Lang	Validity	Fluency	Minimality
en	+12.45	+2.35	−8.34
de	+12.94	−3.17	−6.34
zh	+10.39	+4.61	−4.85
ar	+8.57	+8.34	−5.58
vi	+8.93	−2.81	−2.78
sw	+12.47	−4.46	−37.42
ru	+4.81	−4.90	−3.50

Table 19: Relative degradation (%) under w/o $\mathcal{R}_{\text{aug}}$.

Lang	Validity	Fluency	Minimality
en	−29.01	+54.75	+65.20
de	−32.57	+48.44	+65.93
zh	−32.17	+111.56	+47.60
ar	−30.81	+83.80	+63.19
vi	−28.56	+93.00	+59.66
sw	−54.64	+9.60	+78.83
ru	−33.85	+65.00	+62.64

Table 20: Relative degradation (%) under w/o $\mathcal{R}_{\text{edit}}$.

Lang	$\overline{\lvert\Delta_{\text{val}}\rvert}$	$\overline{\lvert\Delta_{\text{flu}}\rvert}$	$\overline{\lvert\Delta_{\text{min}}\rvert}$	Overall
sw	25.00	5.39	50.84	27.07
zh	19.00	39.92	18.71	25.87
ar	16.50	34.73	24.81	25.35
vi	14.64	33.12	22.40	23.39
en	20.21	20.15	26.80	22.39
de	19.93	18.41	25.60	21.31
ru	15.95	24.00	23.20	21.05

Table 21: Average absolute relative change (%) across all three ablation settings, and overall sensitivity (mean of three dimensions). Higher values indicate greater sensitivity to component removal.

edit constraint. However, this comes at severe cost to *fluency* and *minimality*. Chinese suffers the worst fluency degradation (+111.56%, i.e., PPL more than doubles), followed by Vietnamese and Arabic. Swahili's fluency degradation is remarkably contained (+9.60%), likely because its baseline PPL is already high due to limited pre-training coverage, leaving less room for relative worsening. *Minimality* degrades dramatically across all languages, with Swahili showing the worst (+78.83%), followed by German and English. Chinese has the lowest minimality degradation (+47.60%) under this setting.

M.3.2 Overall Language Sensitivity

Swahili emerges as the most sensitive language overall, in particular its extreme minimality and validity sensitivity. This oscillation between extremes, unique to Swahili, likely stems from its limited pre-training representation, causing the model's behavior to be highly contingent on the specific reward configuration. In contrast, Swahili's fluency sensitivity is the lowest (5.39%), indicating that its generation quality is relatively stable—possibly because the baseline PPL is already high and leaves little room for relative change. **Chinese** and **Arabic** rank second and third overall, but their sensitivity profiles are qualitatively different from Swahili's. Both are dominated by fluency sensitivity. **English** and **German** display the most balanced sensitivity profiles, with all three dimensions falling within a relatively narrow range. English has the highest validity sensitivity among high-resource languages, confirming that its label-flip success depends most critically on the collaboration among all three rewards. German tracks closely behind English in all dimensions. **Russian** is the least sensitive language overall (21.05%), with the lowest validity sensitivity (15.95%). This suggests that Russian counterfactuals are inherently close to the decision boundary and relatively well-formed, making each individual reward component less critical. **Vietnamese** occupies the middle ground, with elevated fluency sensitivity (33.12%) comparable to Chinese and Arabic, but more moderate validity and minimality profiles.

Prediction Prompt — SIB200

```
You are a multilingual topic classifier.
Output exactly one topic label with no additional text.
Classify the primary topic of this {language} text.
Available topics: entertainment, geography, health, politics, science/technology, sports,
travel
Rules:
    • Choose the single best topic the text is mainly about
    • If multiple topics appear, pick the primary subject
    • Output only the label word
Example:
Text: The prime minister announced new election dates.
Answer: politics
Now classify:
Text: {text}
Answer:
```

(a) SIB200

Prediction Prompt — TAXI1500

```
You are a multilingual verse classifier.
Output exactly one label with no additional text.
Classify the category of this {language} verse/text.
Available labels: recommendation, faith, violence, grace, sin, description
Rules:
    • Choose the single best label
    • Output only the label word exactly as provided
Example:
Text: Love your neighbor as yourself.
Answer: recommendation
Now classify:
Text: {text}
Answer:
```

(b) TAXI1500

Figure 8: Prediction prompts used for the two evaluation datasets: SIB200 and TAXI1500. Since the tasks differ across datasets, we adopt dataset-specific prompts for label prediction.

Counterfactual Prompt — SIB200

```
Given a sentence in {language} classified as belonging to one of the topics:
"science/technology", "travel", "politics", "sports", "health", "entertainment", "geography".
Modify the sentence to change its topic to the specified target topic and output the edited
sentence surrounded by <edit>[sentence]</edit> in {language}.

#####Begin Example#####
Original topic: sports
Sentence: The athlete set a new record in the marathon.
Target topic: health
Step 1: Identify key phrases or words determining the original topic: 'athlete', 'record',
'marathon'.
Step 2: Modify these key phrases or words to reflect the target topic (health): 'athlete' to
'patient', 'set a new record' to 'showed improvement', 'marathon' to 'rehabilitation'.
Step 3: Replace the identified words or phrases in the original sentence.
Edited sentence: <edit>The patient showed improvement in the rehabilitation.</edit>
#####End Example####

Request: Given a sentence in {language} classified as belonging to one of the topics:
"science/technology", "travel", "politics", "sports", "health", "entertainment", "geography".
Modify the sentence to change its topic to the specified target topic and output the edited
sentence surrounded by <edit>[sentence]</edit> in {language}.
Original topic: {prediction_label}
Sentence: {text}
Target topic: {target_label}
Edited sentence:
```

(a) SIB200

Counterfactual Prompt — TAXI1500

```
Given a verse/text in {language} classified as belonging to one of the labels: "recommendation",
"faith", "violence", "grace", "sin", "description". Modify the text to change its label to the
specified target label and output ONLY the edited text surrounded by <edit>[text]</edit> in
{language}.

####Begin Example####
Original label: faith
Text: Trust in God and do not be afraid.
Target label: recommendation
Step 1: Identify key phrases or words determining the original label: 'Trust in God', 'do not be
afraid'
Step 2: Modify these phrases to reflect the target label (recommendation): Replace with
advice/actionable guidance.
Step 3: Replace the identified parts in the original text.
Edited text: <edit>Be kind to others and forgive those who wrong you.</edit>
#####End Example####

Request: Modify the following text in {language} to achieve the target label from the original
one.
Original label: {prediction_label}
Text: {text}
Target label: {target_label}
Edited text:
```

(b) TAXI1500

Figure 9: Counterfactual generation prompts used for SIB200 and TAXI1500. For each dataset, both DG-CF and MACRO use the same generation prompt.

SIB200 Dataset Examples

Language	Sentence	Label
ar	يقول الباحثون الرئيسيون أن هذا قد يؤدي إلى الكشف المبكر للسرطان والسل وفيروس العوز المناعي البشري والملاريا في البلدان ذوات الدخل المنخفض، والتي تكون فيها معدلات النجاة من الأمراض مثل سرطان الصدر نصف معدلات البلدان الأغنى.	health
de	Führende Forscher sagen, dass dies die Früherkennung von Krebs, Tuberkulose, HIV und Malaria für Patienten in einkommensschwachen Ländern fördern könnte, wo die Überlebensraten bei Krankheiten wie Brustkrebs teilweise nur halb so hoch sind wie in reicheren Ländern.	health
en	Lead researchers say this may bring early detection of cancer, tuberculosis, HIV and malaria to patients in low-income countries, where the survival rates for illnesses such as breast cancer can be half those of richer countries.	health
ru	Ведущие исследователи утверждают, что он может помочь в раннем выявлении рака, туберкулеза, ВИЧ и малярии у пациентов в странах с низким уровнем дохода, где показатели выживаемости при таких болезнях, как рак молочной железы, могут быть в два раза ниже, чем в более богатых странах.	health
sw	Watafiti wakuu wanasema hili linaweza kuleta ugunduzi wa mapema wa saratani, kifua kikuu, ukimwi na malaria kwa wagonjwa katika nchi zenye mapato ya chini, ambapo viwango vya kupona kwa magonjwa kama vile saratani ya matiti vinaweza kuwa nusu ya vile vya nchi tajiri.	health
vi	Các nhà nghiên cứu chính nói rằng điều này có thể giúp phát hiện sớm bệnh ung thư, bệnh lao, HIV và bệnh sốt rét cho bệnh nhân ở các nước có thu nhập thấp, nơi mà tỷ lệ sống sót khi mắc phải những bệnh như ung thư vú có thể chỉ bằng một nửa tỷ lệ đó ở những nước giàu.	health
zh	主要研究人员表示，这可以让低收入国家/地区的患者尽早发现癌症、肺结核、艾滋病和疟疾。在这些国家/地区，乳腺癌等疾病的生存率可能仅为富裕国家的一半。	health

(a) Dataset Examples of SIB200.

TAXI1500 Dataset Examples

Language	Sentence	Label
ar	وعبد الرب لا يجب ان يخاصم بل يكون مترفقا بالجميع صالحا للتعليم صبورا على المشقات	grace
de	Ein Knechtdes Herrn aber soll nicht streiten, sondern gegen alle milde sein, lehrfähig, duldsam,	grace
en	For a slave of the Lord does not need to fight, but needs to be gentle toward all, qualified to teach, showing restraint when wronged,	grace
ru	рабу же Господа не должно ссориться, но быть приветливым ко всем, учительным, незлобивым,	grace
sw	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mwema kwa kila mtu, awezaye kufundisha, tena mvumilivu.	grace
vi	Vả, tôi tớ của Chúa không nên ưa sự tranh cạnh; nhưng phải ở tử tế với mọi người, có tài dạy dỗ, nhịn nhục,	grace
zh	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心忍耐，	grace

(b) Dataset Examples of TAXI1500.

Figure 10: Dataset examples.

Lang.	Source	Sentence	Prediction
ar	Original	وعبد الرب لا يجب ان يخاصم بل يكون مترفقا بالجميع صالحا للتعليم صبورا على المشقات	grace
ar	DG-CF	وقد يعاقب العبد الرب ويشتد عليه في الاعتراف بالذنوب ويصلح الكفارة	faith
ar	TB-CF	لأن عبد الله لا يحتاج إلى القتال، بل يحتاج إلى أن يكون لطيفًا مع جميع الناس، مؤهلًا للتعليم، ويظهر التحكم في نفسه عند الظلم، ويتجنب الغضب والغرور.	recommendation
ar	MACRO	وعبد الرب لا يجب ان يخاصم بل يكون ظالما للجميع متطاولا على المشقات	sin
de	Original	Ein Knechtdes Herrn aber soll nicht streiten, sondern gegen alle milde sein, lehrfähig, duldsam,	grace
de	DG-CF	Ein Knechtdes Herrn soll streiten und nicht mild sein, sondern zornig, unlehrend und ungnädig.	faith
de	TB-CF	Ein Knecht des Herrn muss nicht kämpfen, sondern muss sanftmütig gegenüber allen sein, fähig, zu lehren, sich bei Unrecht beherrschen und Zorn und Stolz vermeiden.	grace
de	MACRO	Ein Knechtdes Herrn soll streiten und gegen alle harte sein, lehrunfähig, unduldsam.	faith
en	Original	For a slave of the Lord does not need to fight , but needs to be gentle toward all , qualified to teach , showing restraint when wronged ,	grace
en	DG-CF	For a slave of the Lord does not need to fight, but needs to be gentle toward all, qualified to teach, showing restraint when wronged, and avoiding anger and pride.	grace
en	MACRO	For a slave of the Lord does not need to fight, but needs to be cruel toward all, qualified to teach, showing no restraint when wronged,	violence
ru	Original	рабу же Господа не должно ссориться, но быть приветливым ко всем, учительным, незлобивым,	grace
ru	DG-CF	Рабу же Господа должно ссориться, быть грубым к другим, неучительным, злобным.	violence
ru	TB-CF	Рабу Господа не нужно бороться, но нужно быть добрее ко всем, быть подготовленным к преподаванию, проявлять терпение в случае несправедливости и избегать гнева и гордыни.	grace
ru	MACRO	Рабу же Господа должно ссориться, быть неприветливым к некоторым, злобным и неучительным.	faith
sw	Original	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mwema kwa kila mtu, awezaye kufundisha, tena mvumilivu.	grace
sw	DG-CF	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mwema kwa kila mtu, awezaye kufundisha, tena mvumilivu.	grace
sw	TB-CF	Mwanaume wa Mungu hakuna kushambulika, lakini inaweza kushuka kote kote, amekuwa na makini ya kutoa maelezo, amekuwa na makini ya kushuka kama kuna kushuka, na kushuka kama kuna kushuka.	grace
sw	MACRO	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mabadiliko kwa kila mtu, awezaye kuongeza kosa, tena kosa.	faith
vi	Original	Vả, tôi tớ của Chúa không nên ưa sự tranh cạnh; nhưng phải ở tử tế với mọi người, có tài dạy dỗ, nhịn nhục,	grace
vi	DG-CF	Vả, tôi tớ của Chúa không nên ưa sự tranh cạnh; nhưng phải ở tử tế với mọi người, có tài dạy dỗ, và phải biết tức giận kẻ lầm lỗi.	faith
vi	TB-CF	Một người đầy tớ của Chúa không cần phải chiến đấu, nhưng cần phải hiền lành với tất cả mọi người, đủ điều kiện để giảng dạy, biết kiềm chế khi bị hiểu lầm, và tránh sự tức giận và tự phụ.	grace
vi	MACRO	Và tôi tớ của Chúa nên ưa sự tranh cạnh; phải lăng mạ mọi người, không có tài dạy dỗ, kiêu căng.	violence
zh	Original	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心忍耐，	grace
zh	DG-CF	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心忍耐，否则便是行恶。	grace
zh	TB-CF	一个属主的仆人不需要争战，却要对所有人温柔，有资格教导人，在受委屈时能忍耐，避免愤怒和骄傲。	grace
zh	MACRO	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心诡诈。	faith

Table 8: Multilingual counterfactual examples from the TAXI1500 task. The figure shows original Bible verses and their labels together with counterfactual samples generated by three methods (DG-CF, TB-CF, and MACRO) across multiple languages using Qwen3-8B.

Lang.	Source	Sentence	Prediction
ar	Original	"يقول الباحثون الرئيسيون أن هذا قد يؤدي إلى الكشف المبكر للسرطان والسل وفيروس العوز المناعي البشري والملاريا في البلدان ذوات الدخل المنخفض، والتي تكون فيها معدلات النجاة من الأمراض مثل سرطان الصدر نصف معدلات البلدان الأغنى."	health
ar	DG-CF	يقول الباحثون الرئيسيون أن هذا قد يؤدي إلى الكشف المبكر عن الأمراض أثناء السفر في البلدان ذوات الدخل المنخفض، والتي تكون فيها معدلات النجاة من الأمراض مثل سرطان الصدر نصف معدلات البلدان الأغنى.	health
ar	TB-CF	يقول الباحثون الرئيسيون إن هذا قد يجلب الكشف المبكر عن السرطان والسل والفيروسات مثل الإيدز والملاريا للمسافرين في الدول ذات الدخل المنخفض، حيث يمكن أن تكون معدلات البقاء على قيد الحياة من الأمراض مثل سرطان الثدي نصف ما هي في الدول الغنية.	travel
ar	MACRO	يقول الباحثون الرئيسيون أن هذا قد يؤدي إلى الكشف المبكر للسفر والسل وفيروس العوز المناعي البشري والملاريا في البلدان ذوات الدخل المنخفض، والتي تكون فيها معدلات النجاة من الأمراض مثل سرطان الصدر نصف معدلات البلدان الأغنى.	science/technology
de	Original	Führende Forscher sagen, dass dies die Früherkennung von Krebs, Tuberkulose, HIV und Malaria für Patienten in einkommensschwachen Ländern fördern könnte, wo die Überlebensraten bei Krankheiten wie Brustkrebs teilweise nur halb so hoch sind wie in reicheren Ländern.	science/technology
de	DG-CF	Führende Reiseexperten sagen, dass dies die Früherkennung von Krebs, Tuberkulose, HIV und Malaria für Reisende in einkommensschwachen Ländern fördern könnte, wo die Überlebensraten bei Krankheiten wie Brustkrebs teilweise nur halb so hoch sind wie in reicheren Ländern.	travel
de	TB-CF	Führende Forscher sagen, dies könnte eine frühzeitige Erkennung von Krebs, Tuberkulose, HIV und Malaria für Reisende in einkommensschwachen Ländern ermöglichen, wo die Überlebensraten bei Krankheiten wie Brustkrebs halb so hoch sind wie in reicheren Ländern.	science/technology
de	MACRO	Führende Forscher sagen, dass dies die Früherkennung von Reisen, Tourismus, Flughäfen und Museen für Touristen in einkommensschwachen Ländern fördern könnte, wo die Überlebensraten bei Reisezielen wie fremden Kulturen teilweise nur halb so hoch sind wie in reicheren Ländern.	science/technology
en	Original	Lead researchers say this may bring early detection of cancer, tuberculosis, HIV and malaria to patients in low-income countries, where the survival rates for illnesses such as breast cancer can be half those of richer countries.	science/technology
en	DG-CF	Lead researchers say this may bring early detection of cancer, tuberculosis, HIV and malaria to travelers in low-income countries, where the survival rates for illnesses such as breast cancer can be half those of richer countries.	health
en	MACRO	Lead travelers say this may bring early discovery of destinations, tuberculosis, HIV and malaria to tourists in low-income countries, where the survival rates for illnesses such as breast cancer can be half those of richer countries.	health
ru	Original	Ведущие исследователи утверждают, что он может помочь в раннем выявлении рака, туберкулеза, ВИЧ и малярии у пациентов в странах с низким уровнем дохода, где показатели выживаемости при таких болезнях, как рак молочной железы, могут быть в два раза ниже, чем в более богатых странах.	health
ru	DG-CF	Ведущие исследователи утверждают, что он может помочь в раннем выявлении достопримечательностей, традиций и культурных особенностей при путешествиях по странам с низким уровнем дохода, где показатели интереса к местным достопримечательностям могут быть в два раза ниже, чем в более богатых странах.	travel
ru	TB-CF	Ведущие исследователи говорят, что это может позволить раннее выявление рака, туберкулеза, ВИЧ и малярии у путешественников в странах с низким уровнем дохода, где выживаемость при заболеваниях, таких как рак молочной железы, может составлять половину показателей более развитых стран.	health
ru	MACRO	Ведущие исследователи утверждают, что он может помочь в раннем выявлении туристических точек, традиций, достопримечательностей и культурных особенностей в странах с низким уровнем дохода, где показатели интереса к путешествиям, как посещение памятников, могут быть в два раза ниже, чем в более богатых странах.	travel
sw	Original	Watafiti wakuu wanasema hili linaweza kuleta ugunduzi wa mapema wa saratani, kifua kikuu, ukimwi na malaria kwa wagonjwa katika nchi zenye mapato ya chini, ambapo viwango vya kupona kwa magonjwa kama vile saratani ya matiti vinaweza kuwa nusu ya vile vya nchi tajiri.	health
sw	DG-CF	Watafiti wakuu wanasema hili linaweza kuleta ugunduzi wa mapema ya mitihani ya mfulu, mafanikio ya mfulu, ukimwi na malaria kwa wagonjwa katika nchi zenye mapato ya chini, ambapo viwango vya kupona kwa magonjwa kama vile saratani ya matiti vinaweza kuwa nusu ya vile vya nchi tajiri.	health
sw	TB-CF	Wanaresearcher wa kwanza wanaonyesha kuwa hii inaweza kubadilisha ujumla wa kusafiri kwa kuzingatia ujumla wa kusafiri kwa kuzingatia ujumla wa kusafiri kwa kuzingatia ujumla wa kusafiri kwa kuzingatia ujumla wa kusafiri ...	travel
sw	MACRO	Watafiti wakuu wanasema hili linaweza kuleta ugunduzi wa mapema wa saratani, kifua kikuu, ukimwi na malaria kwa wagonjwa katika nchi zenye mapato ya chini, ambapo viwango vya kupona kwa magonjwa kama vile saratani ya matiti vinaweza kuwa nusu ya vile vya nchi tajiri.	health
vi	Original	Các nhà nghiên cứu chính nói rằng điều này có thể giúp phát hiện sớm bệnh ung thư, bệnh lao, HIV và bệnh sốt rét cho bệnh nhân ở các nước có thu nhập thấp, nơi mà tỷ lệ sống sót khi mắc phải những bệnh như ung thư vú có thể chỉ bằng một nửa tỷ lệ đó ở những nước giàu.	health
vi	DG-CF	Các nhà nghiên cứu chính nói rằng điều này có thể giúp phát hiện sớm các điểm đến du lịch tiềm năng, bệnh lao, HIV và bệnh sốt rét cho du khách ở các nước có thu nhập thấp, nơi mà tỷ lệ sống sót khi mắc phải những bệnh như ung thư vú có thể chỉ bằng một nửa tỷ lệ đó ở những nước giàu.	travel
vi	TB-CF	Những nhà nghiên cứu hàng đầu cho biết điều này có thể giúp phát hiện sớm các bệnh như ung thư, lao, HIV và sốt rét ở du khách tại các quốc gia thu nhập thấp, nơi tỷ lệ sống sót đối với các bệnh như ung thư vú có thể chỉ bằng một nửa so với ở các nước giàu hơn.	health
vi	MACRO	Các nhà nghiên cứu chính nói rằng điều này có thể giúp phát hiện sớm các điểm du lịch, bệnh lao, HIV và bệnh sốt rét cho bệnh nhân ở các nước có thu nhập thấp, nơi mà tỷ lệ sống sót khi mắc phải những điểm du lịch như ung thư vú có thể chỉ bằng một nửa tỷ lệ đó ở những nước giàu.	health
zh	Original	主要研究人员表示，这可以让低收入国家/地区的患者尽早发现癌症、肺结核、艾滋病和疟疾。在这些国家/地区，乳腺癌等疾病的生存率可能仅为富裕国家的一半。	health
zh	DG-CF	主要旅行专家表示，这可以让游客在低收入国家/地区更早发现癌症、肺结核、艾滋病和疟疾。在这些国家/地区，乳腺癌等疾病的生存率可能仅为富裕国家的一半。	travel
zh	TB-CF	首席研究人员表示，这可能为前往低收入国家的旅行者带来早期发现癌症、结核病、艾滋病和疟疾的能力，在这些国家，像乳腺癌这样的疾病存活率可能仅为高收入国家的一半。	science/technology
zh	MACRO	主要旅游研究人员表示，这可以让低收入国家/地区的游客尽早发现癌症、肺结核、艾滋病和疟疾。在这些国家/地区，乳腺癌等疾病的生存率可能仅为富裕国家的一半。	travel

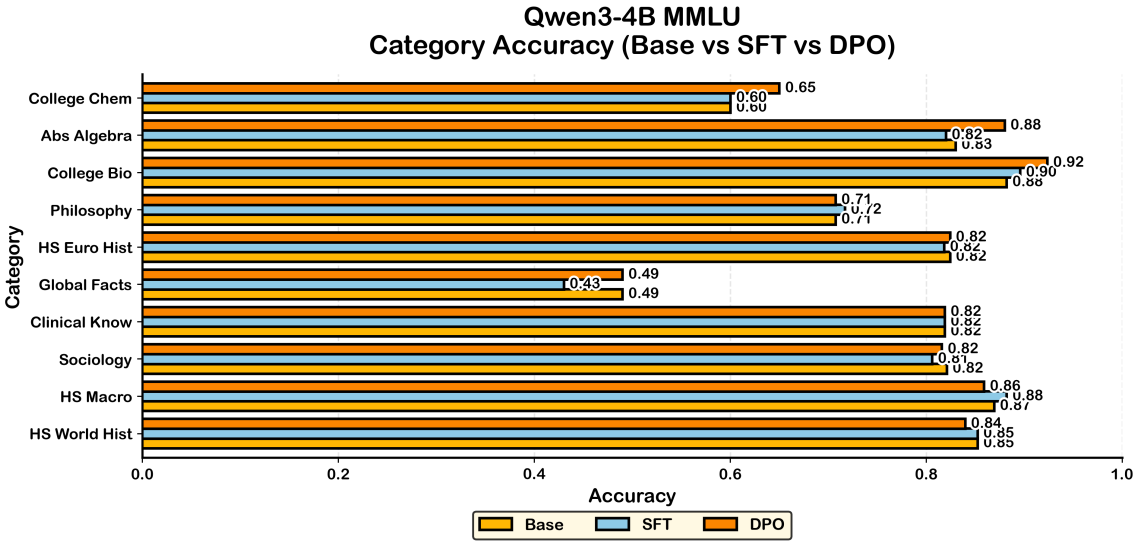
Table 9: Multilingual counterfactual examples from the SIB200 task. The figure shows original sentences and their labels together with counterfactual samples generated by three methods (DG-CF, TB-CF, and MACRO) across multiple languages using Qwen3-8B.

Lang.	Source	Sentence	Prediction
ar	Original	تضيف الطفرة اختلافات وراثيّة جديدة، ويزيلها الانتقاء من تجميعة الاختلافات التعبيريّة.	science/technology
ar	MACRO	تضيف السياسة اختلافات وراثيّة جديدة، ويزيلها الانتقاء من تجميعة الاختلافات التعبيريّة.	science/technology
ar	w/o $\mathcal{R}_{flip}$	تضيف السياسة اختلافات وراثيّة جديدة، ويزيلها الانتقاء من تجميعة الاختلافات التعبيريّة.	science/technology
ar	w/o $\mathcal{R}_{aug}$	تضيف السياسة اختلافات وراثيّة جديدة، ويزيلها التفاهم من تجميعة الاختلافات التعبيريّة.	politics
ar	w/o $\mathcal{R}_{edit}$	تضيف المفاوضات اختلافات سياسية جديدة، ويزيلها الحوار من تجميعة الاختلافات بين الدول.	politics
de	Original	Durch den Wechsel vom Viertelmeilen- zum Halbmeilen-Rennen ist das Tempo weitaus weniger wichtig und Ausdauer eine absolute Notwendigkeit.	sports
de	MACRO	Durch den Wechsel vom Regionalwahl- zum Nationalwahl-System ist die Parteizugehörigkeit weitaus weniger wichtig und Koalitionsbildung eine absolute Notwendigkeit.	politics
de	w/o $\mathcal{R}_{flip}$	Durch den Wechsel vom Wahlkreis zum Bundestag ist die Parteizugehörigkeit weitaus weniger wichtig und die Erfahrung eine absolute Notwendigkeit.	politics
de	w/o $\mathcal{R}_{aug}$	Durch den Wechsel vom Parteienwechsel zum Regierungsbildungsprozess ist die Parteispaltung weitaus weniger wichtig und Koalitionsfähigkeit eine absolute Notwendigkeit.	politics
de	w/o $\mathcal{R}_{edit}$	Durch den Wechsel von einer Parteienkoalition zur Regierungsbildung ist die Meinungsbildung weitaus weniger wichtig und die politische Erfahrung eine absolute Notwendigkeit.	politics
en	Original	With the change from the quarter to the half mile run, speed becomes of much less importance and endurance becomes an absolute necessity.	sports
en	MACRO	With the change from the policy to the legislative agenda, strategy becomes of much less importance and compromise becomes an absolute necessity.	politics
en	w/o $\mathcal{R}_{flip}$	With the change from the policy to the legislative mile run, speed becomes of much less importance and endurance becomes an absolute necessity.	science/technology
en	w/o $\mathcal{R}_{aug}$	With the change from the policy to the legislative agenda, power becomes of much less importance and compromise becomes an absolute necessity.	science/technology
en	w/o $\mathcal{R}_{edit}$	With the shift from one political party to another, ideology becomes of much less importance and compromise becomes an absolute necessity.	politics
ru	Original	Если увеличить расстояние для бега с четверти до половины мили, скорость становится не так важна, тогда как выносливость превращается в абсолютную необходимость.	sports
ru	MACRO	Если увеличить расстояние для дебатов с четверти до половины часа, скорость становится не так важна, тогда как умение говорить превращается в абсолютную необходимость.	science/technology
ru	w/o $\mathcal{R}_{flip}$	Если увеличить расстояние для дебатов с четверти до половины часа, скорость становится не так важна, тогда как выносливость превращается в абсолютную необходимость.	sports
ru	w/o $\mathcal{R}_{aug}$	Если увеличить расстояние для дебатов с четверти до половины часа, скорость становится не так важна, тогда как убеждённость превращается в абсолютную необходимость.	science/technology
ru	w/o $\mathcal{R}_{edit}$	Если увеличить срок полномочий политика с четырех до пяти лет, стабильность становится не так важна, тогда как доверие превращается в абсолютную необходимость.	politics
sw	Original	Kwa mabadiliko kutoka mbio za robo hadi nusu maili, kasi huwa yenye umuhimu mdogo zaidi na ustahimilivu ukawa hitaji lenye umuhimu kabisa.	science/technology
sw	MACRO	Kwa mabadiliko kutoka mbio za kura hadi nusu maili, kasi huwa yenye umuhimu mdogo zaidi na ustahimilivu ukawa hitaji lenye umuhimu kabisa.	science/technology
sw	w/o $\mathcal{R}_{flip}$	Kwa mabadiliko kutoka mbio za pambana hadi nusu maili, kasi huwa yenye umuhimu mdogo zaidi na ustahimilivu ukawa hitaji lenye umuhimu kabisa.	science/technology
sw	w/o $\mathcal{R}_{aug}$	Kwa mabadiliko kutoka mbio za ujanja hadi nusu maili, kasi huwa yenye mafundisho mdogo zaidi na ustahimilivu ujanja hitaji lenye mafundisho kabisa.	science/technology
sw	w/o $\mathcal{R}_{edit}$	Kwa mabadiliko kutoka majukumu ya wale wanaume hadi kauli ya miji, mamlaki huwa yenye umuhimu mdogo zaidi na uhanika wa mamlaki hitaji lenye umuhimu kabisa.	health
vi	Original	Với sự thay đổi từ 1/4 sang 1/2 dặm chạy, tốc độ trở nên ít quan trọng hơn nhiều và sức bền trở thành điều kiện tiên quyết.	science/technology
vi	MACRO	Với sự thay đổi từ 1/4 sang 1/2 phiếu bầu, số phiếu trở nên ít quan trọng hơn nhiều và sự đồng thuận trở thành điều kiện tiên quyết.	politics
vi	w/o $\mathcal{R}_{flip}$	Với sự thay đổi từ 1/4 sang 1/2 phiếu biểu quyết, tốc độ trở nên ít quan trọng hơn nhiều và sự đồng thuận trở thành điều kiện tiên quyết.	science/technolog
vi	w/o $\mathcal{R}_{aug}$	Với sự thay đổi từ 1/4 sang 1/2 phiếu bầu, tốc độ trở nên ít quan trọng hơn nhiều và sự nhất trí trở thành điều kiện tiên quyết.	politics
vi	w/o $\mathcal{R}_{edit}$	Với sự thay đổi từ một chính sách mới sang một hiệp ước quốc tế, quyền lực trở nên ít quan trọng hơn nhiều và hợp tác trở thành điều kiện tiên quyết.	politics
zh	Original	突变会添加新的基因变异，而选择则会从表达的基因变异池中把变异移除。	science/technology
zh	MACRO	突变会添加新的政策变动，而选择则会从表达的政策变动池中把变动移除。	politics
zh	w/o $\mathcal{R}_{flip}$	突变会添加新的政策变异，而选择则会从表达的政策变异池中把变异移除。	science/technology
zh	w/o $\mathcal{R}_{aug}$	突变会添加新的政治变异，而选择则会从表达的政治变异池中把变异移除。	politics
zh	w/o $\mathcal{R}_{edit}$	政策变化会引入新的法规调整，而选举则会从现行的政策体系中淘汰某些规定。	politics

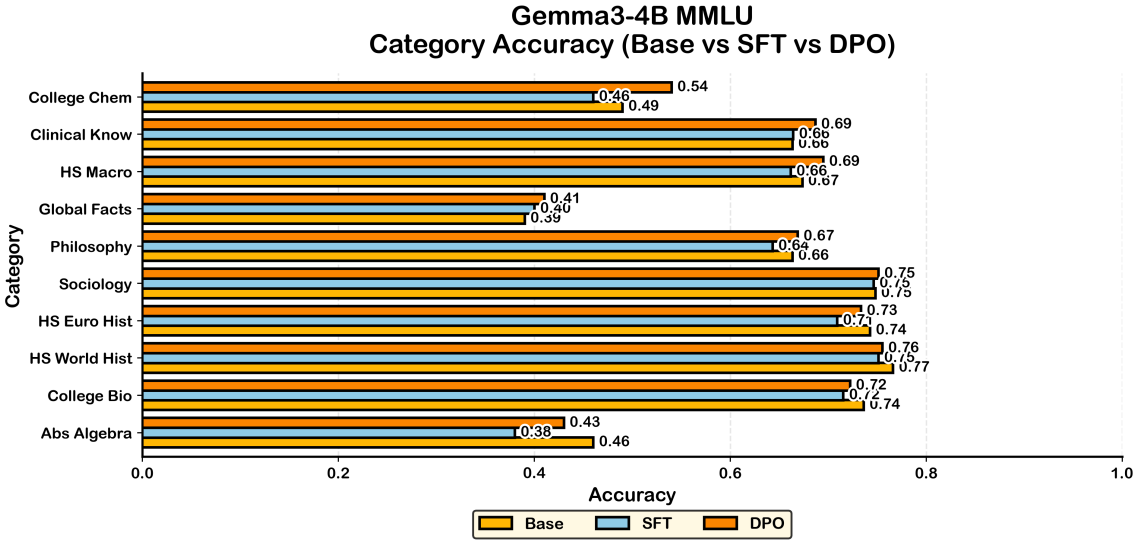
Table 10: Multilingual counterfactual examples from the SIB200 task. The figure shows original sentences and their labels together with counterfactual samples generated by three methods (MACRO and its different reward ablations) across multiple languages using Qwen3-8B.

Lang.	Source	Sentence	Prediction
ar	Original	وعبد الرب لا يجب ان يخاصم بل يكون مترفقا بالجميع صالحا للتعليم صبورا على المشقات	grace
ar	MACRO	وعبد الرب لا يجب ان يخاصم بل يكون ظالما للجميع متطاولا على المشقات	sin
ar	w/o $\mathcal{R}_{flip}$	وأصبح العبد ينكر أمر الله ويحمل على القيم منتهكا للحدود	sin
ar	w/o $\mathcal{R}_{aug}$	>وتبعل الرب لا يجب أن يظلم بل يكون مؤذيا للجميع منكرًا للتعليم جبانًا على الأمور<	violence
ar	w/o $\mathcal{R}_{edit}$	وَمَن يَسْتَكْبِرُ وَيَتَّخِذَ الْعِبَادَ أَصْغَرَ مِنْهُ فَإِنَّهُ أَبْدَى فِي الْعَذَابِ	violence
de	Original	Ein Knechtdes Herrn aber soll nicht streiten, sondern gegen alle milde sein, lehrfähig, duldsam,	grace
de	MACRO	Ein Knechtdes Herrn soll streiten und gegen alle harte sein, lehrunfähig, unduldsam.	faith
de	w/o $\mathcal{R}_{flip}$	Ein Knechtdes Herrn soll streiten, nicht milde sein, und gegenüber allen unhöflich handeln.	violence
de	w/o $\mathcal{R}_{aug}$	Ein Knechtdes Herrn aber soll streiten, gewalttätig sein, unerträglich und boshaft.	violence
de	w/o $\mathcal{R}_{edit}$	Ein Knechtdes Herrn soll streiten und allen gegenüber unhöflich sein, unweise, ungeduldig und böse.	violence
en	Original	For a slave of the Lord does not need to fight , but needs to be gentle toward all , qualified to teach , showing restraint when wronged ,	grace
en	MACRO	For a slave of the Lord does not need to fight, but needs to be cruel toward all, qualified to teach, showing no restraint when wronged,	violence
en	w/o $\mathcal{R}_{flip}$	For a slave of the Lord does not need to fight, but needs to be cruel toward all, qualified to deceive, showing no restraint when wronged,	violence
en	w/o $\mathcal{R}_{aug}$	A slave of the Lord must fight, and should not be gentle toward all, unqualified to teach, and show no restraint when wronged.	grace
en	w/o $\mathcal{R}_{edit}$	Being cruel and unkind to others is a sin, and those who reject righteousness are condemned.	sin
ru	Original	рабу же Господа не должно ссориться, но быть приветливым ко всем, учительным, незлобивым,	grace
ru	MACRO	Рабу же Господа должно ссориться, быть неприветливым к некоторым, злобным и неучительным.	faith
ru	w/o $\mathcal{R}_{flip}$	Рабу же Господа должно ссориться, быть грубым ко всем, злобным и непримиримым.	violence
ru	w/o $\mathcal{R}_{aug}$	Рабу же Господа должно ссориться, быть грубым ко всем, злобным и кричащим.	violence
ru	w/o $\mathcal{R}_{edit}$	Рабу же Господа должно ссориться, быть злобным, грубым и непримиримым к другим.	violence
sw	Original	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mwema kwa kila mtu, awezaye kufundisha, tena mvumilivu.	grace
sw	MACRO	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mabadiliko kwa kila mtu, awezaye kuongeza kosa, tena kosa.	faith
sw	w/o $\mathcal{R}_{flip}$	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mwingi kwa kila mtu, awezaye kufundisha, tena mvumilivu.	grace
sw	w/o $\mathcal{R}_{aug}$	Tena haimpasi mtumishi wa Bwana kuwa mgomvi, bali inampasa awe mtu mwingi kwa kila mtu, awezaye kufundisha, tena mvumilivu.	grace
sw	w/o $\mathcal{R}_{edit}$	Kuwa mtu wa kikosi, kuongea kwa mizigo, kumwonya mtu binafsi, na kuwa mpona wa mitume na kifungua.	grace
vi	Original	Và, tôi tớ của Chúa không nên ưa sự tranh cạnh; nhưng phải ở tử tế với mọi người, có tài dạy dỗ, nhịn nhục,	grace
vi	MACRO	Và tôi tớ của Chúa nên ưa sự tranh cạnh; phải lăng mạ mọi người, không có tài dạy dỗ, kiêu căng.	violence
vi	w/o $\mathcal{R}_{flip}$	Và, tôi tớ của Chúa nên ưa sự tranh cạnh; nhưng phải sống ích kỷ với mọi người, không có tài dạy dỗ, không nhịn nhục.	grace
vi	w/o $\mathcal{R}_{aug}$	Và, tôi tớ của Chúa nên thích đấu tranh, không phải ở tử tế với mọi người, không có tài dạy dỗ, không nhịn nhục.	violence
vi	w/o $\mathcal{R}_{edit}$	Hãy sống trong sự gian ác, không biết tôn trọng người khác và không ngại làm điều sai trái.	violence
zh	Original	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心忍耐，	grace
zh	MACRO	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心诡诈。	faith
zh	w/o $\mathcal{R}_{flip}$	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心忍耐，但不可行恶。	grace
zh	w/o $\mathcal{R}_{aug}$	然而主的仆人不可争竞，只要温温和和地待众人，善于教导，存心忍耐，却要行恶。	violence
zh	w/o $\mathcal{R}_{edit}$	凡行奸淫的，必遭神的审判；凡说谎的，必被永远定罪。	sin

Table 11: Multilingual counterfactual examples from the TAXI1500 task. The figure shows original sentences and their labels together with counterfactual samples generated by three methods (MACRO and its different reward ablations) across multiple languages using Qwen3-8B.

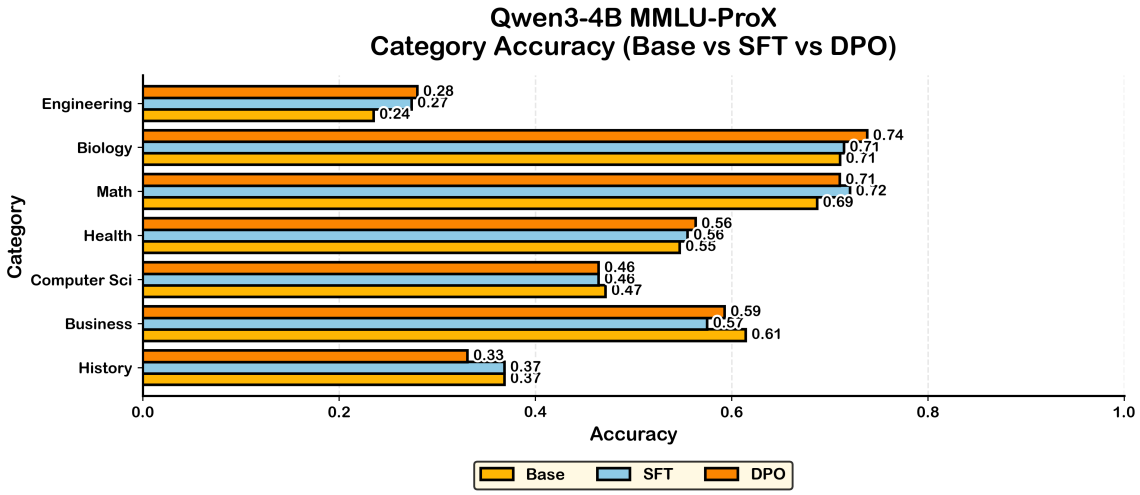


(a) Category-wise performance on MMLU with Qwen3-4B.

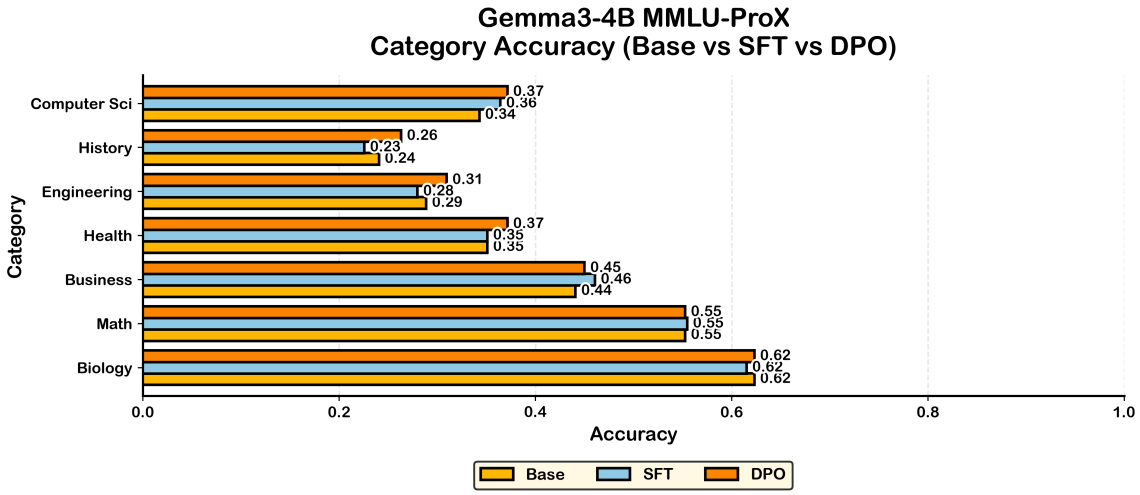


(b) Category-wise performance on MMLU with Gemma3-4B.

Figure 11: Impact of MACRO on multilingual general capability measured on MMLU.

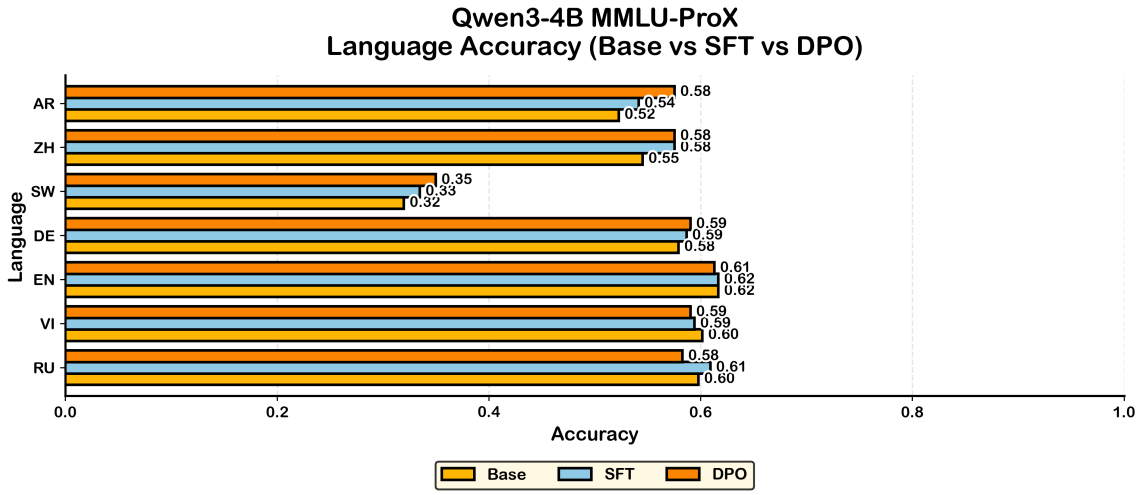


(a) Category-wise performance aggregated over the languages used in post-training on MMLU-ProX with `Qwen3-4B`.

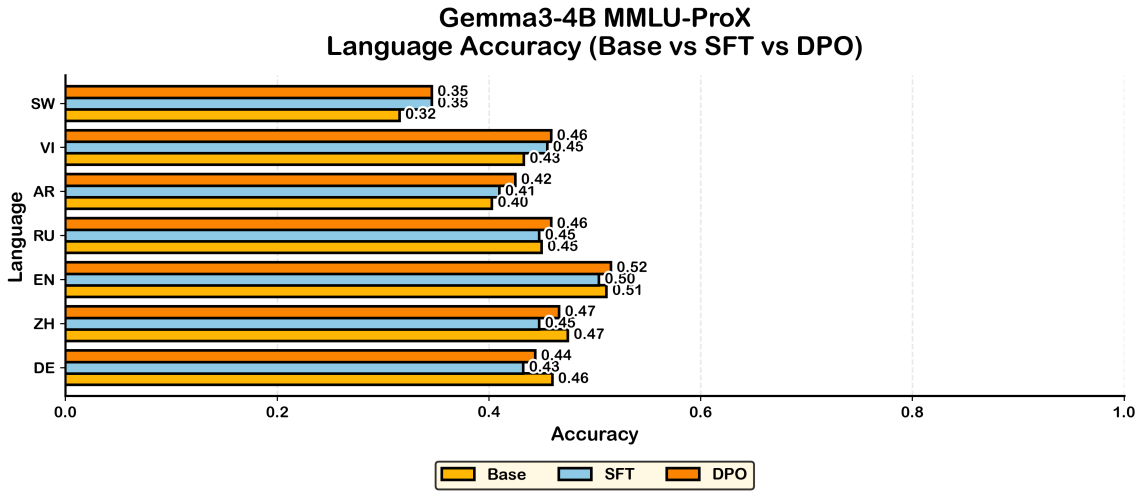


(b) Category-wise performance aggregated over the languages used in post-training on MMLU-ProX with `Gemma3-4B`.

Figure 12: Impact of MACRO on reasoning capability measured on MMLU-ProX from the category perspective . Subfigures (a) and (b) present the category-wise performance of `Qwen3-4B` and `Gemma3-4B`, respectively.



(a) Language-wise performance on MMLU-ProX with Qwen3-4B.



(b) Language-wise performance on MMLU-ProX with Gemma3-4B.

Figure 13: Impact of MACRO on cross-lingual generalization measured on MMLU-ProX from the language perspective. Subfigures (a) and (b) present the language-wise performance of Qwen3-4B and Gemma3-4B, respectively.

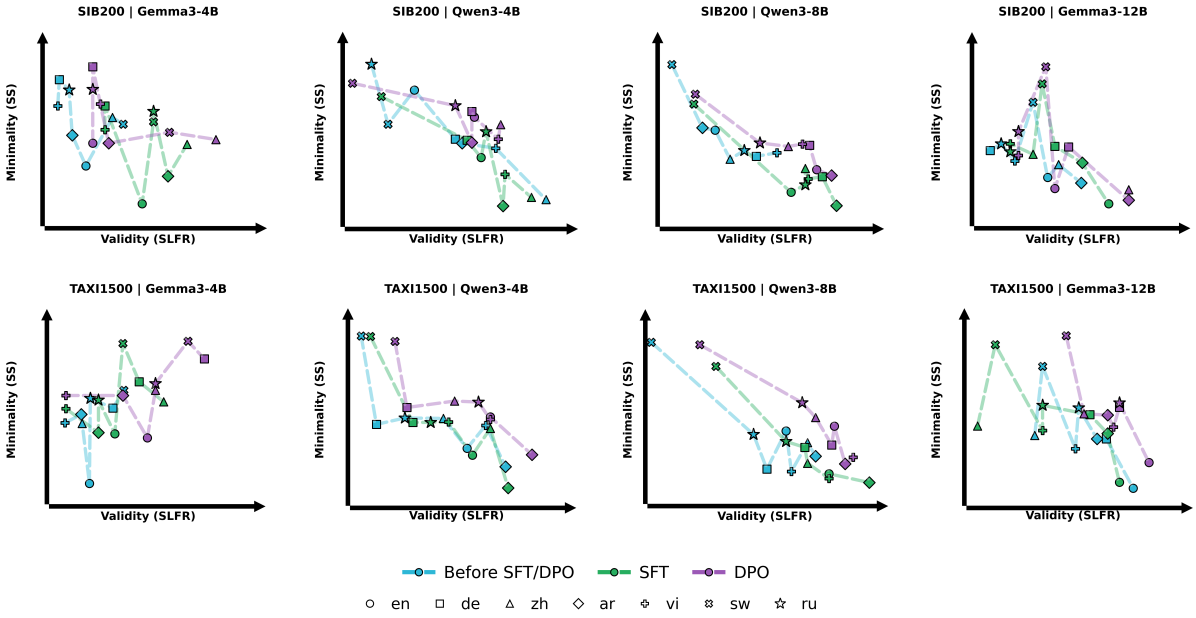


Figure 14: The validity-minimality trade-off across languages across all models on SIB200 and Taxi1500.

(a) Cross-lingual Edit Similarity on TAXI1500, computed from counterfactuals in different languages.

(b) Cross-lingual Edit Similarity on SIB200, computed from counterfactuals in different languages.

Figure 15: Cross-lingual edit similarity scores.